

Working with Gene Lists and Over-representation Analysis

MBV-INF4410

Tuesday, September 15th, 2009

Ian Donaldson

<http://donaldson.uio.no>

This talk is a remix of two talks presented in 2009 at the Canadian Bioinformatics Workshops by Gary Bader and Quaid Morris. Many thanks to Gary, Quaid and the CBW for making this material available.

Ian Donaldson, September 14th

<http://baderlab.org>

<http://morrislab.med.utoronto.ca/>

<http://www.bioinformatics.ca/workshops/2009/course-content>

(see Interpreting Gene Lists from -omics Studies.

Module 1: Introduction to gene lists (Chair: Gary Bader) and

Module 2: Finding over-represented gene functions in gene lists (Chair: Quaid Morris)

This page is available in the following languages:

Afrikaans বাংলাৰাখী Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.

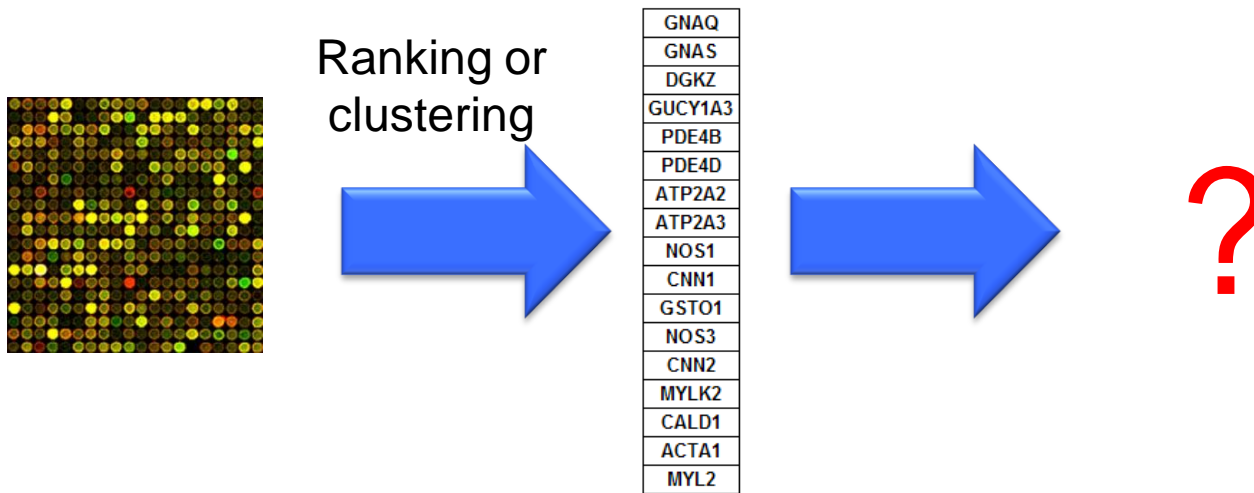
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
[English](#) [French](#)

Gene Lists Overview

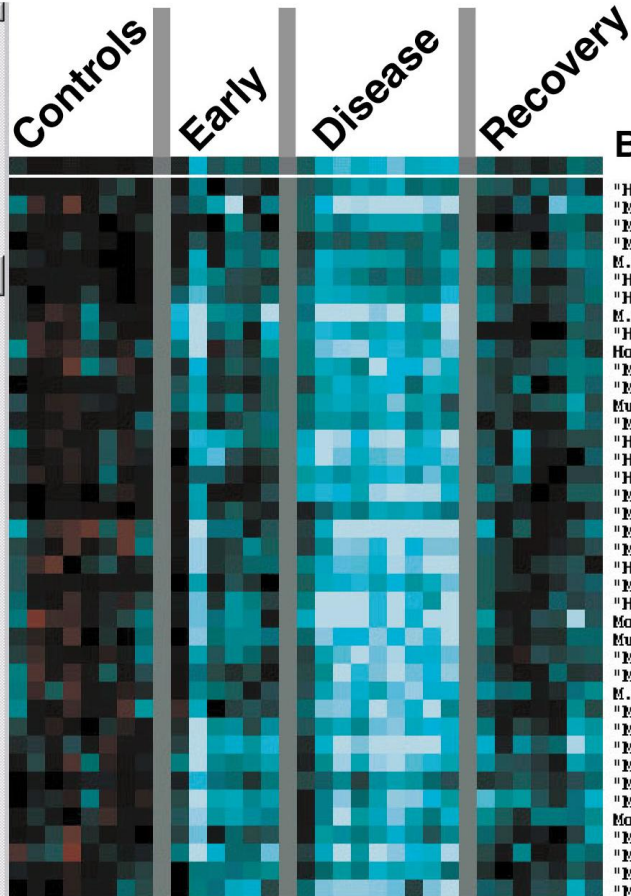
- Interpreting gene lists
- Gene attributes
 - Gene Ontology
 - Ontology Structure
 - Annotation
 - BioMart + other sources
- Gene identifiers and mapping
- Part 2: Network Introduction

Interpreting Gene Lists

- My cool new screen worked and produced 1000 hits! ...Now what?
- Genome-Scale Analysis (Omics)
 - Genomics, Proteomics



Cardiomyopathy: Downregulated Genes

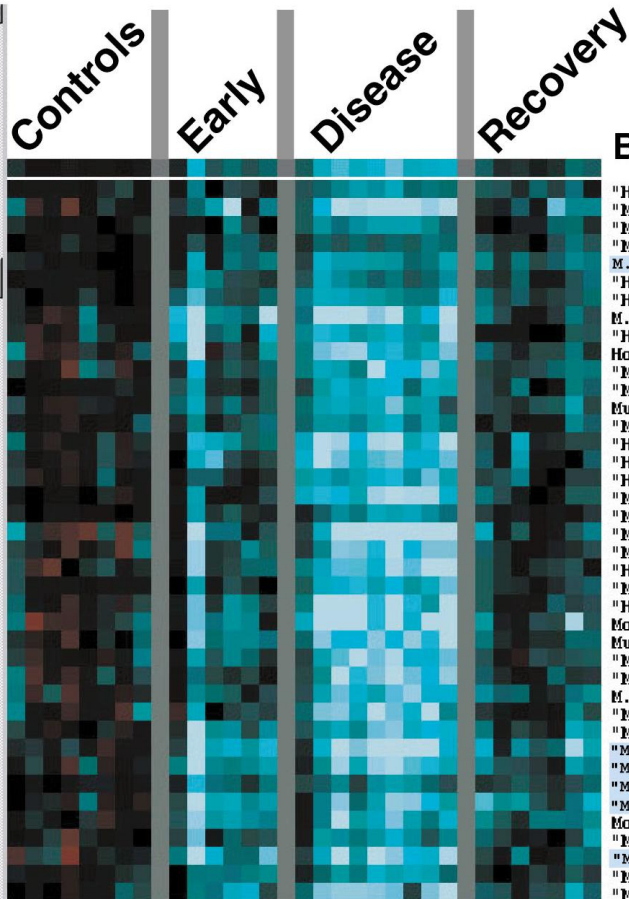


BLAST Definitions

"Homologous to sn 007021: PRE-mRNA SPLICING FACTOR SF2, P32 SUBUNIT PRECURSOR (GCIQ-R PROTEIN)
 "Mus musculus proteasome activator PA28 alpha subunit mRNA, complete cds"
 "Mus musculus cdc37 homolog mRNA, complete cds"
 "Mus musculus ornithine decarboxylase antisense gene, complete cds"
 M.musculus mRNA for carnitine acetyltransferase
 "Homologous to sn 000779: CALCIUM-TRANSPORTING ATPASE SARCOPLASMIC RETICULUM TYPE (EC 3.6.1.34)
 "Homologous to sn P11507: CALCIUM-TRANSPORTING ATPASE ENDOPLASMIC RETICULUM TYPE (EC 3.6.1.38)
 M.musculus ENO3 mRNA for enolase beta subunit
 "Homologous to sn P47858: 6-PHOSPHOFRUCTOKINASE, MUSCLE TYPE (EC 2.7.1.11) (PHOSPHOFRUCTOKINASE)
 Homologous to sn P23327: SARCOPLASMIC RETICULUM HISTIDINE-RICH CALCIUM-BINDING PROTEIN PRECURSOR
 "Mouse AE3 mRNA, complete cds"
 "M.musculus glucose transporter 2 mRNA, complete cds"
 Mus musculus aspartate aminotransferase gene 5'-flank and exon 1
 "Mus musculus thioredoxin-dependent peroxide reductase (tox) mRNA, complete cds"
 "Homologous to sn P47858: 6-PHOSPHOFRUCTOKINASE, MUSCLE TYPE (EC 2.7.1.11) (PHOSPHOFRUCTOKINASE)
 "Homologous to sn P11508: CALCIUM-TRANSPORTING ATPASE SARCOPLASMIC RETICULUM TYPE (EC 3.6.1.34)."
 "Homologous to sn P35434: ATP SYNTHASE DELTA CHAIN, MITOCHONDRIAL PRECURSOR (EC 3.6.1.34)."
 "Mus musculus F1FOATP synthase complex E subunit (Atp5k) gene, complete cds"
 "Mus musculus NAD(H)-specific isocitrate dehydrogenase gamma subunit precursor, mRNA, complete cds"
 "M.musculus gene for dodecenoyl-CoA delta-isomerase, exons 1 and 2"
 "Mus musculus cytochrome c oxidase subunit VIII-H precursor (COX8H) mRNA, complete cds"
 "Homologous to sn P35745: ACYLPHOSPHATASE, MUSCLE TYPE ISOZYME (EC 3.6.1.7) (ACYLPHOSPHATE PHOSPHATASE)
 "Mus musculus CD-1 cardiac troponin I mRNA, complete cds"
 "Homologous to sn P00566: CREATINE KINASE, M CHAIN (EC 2.7.3.2) (MU-2 PROTEIN)."
 Mouse mRNA for protein with homology to transition protein 2 (TP2)
 Mus musculus Selenium-binding liver protein mRNA
 "Mus musculus (clone MAR1) aldose reductase mRNA, complete cds"
 "Mus musculus vascular endothelial growth factor B 186 (VEGF-B) precursor, mRNA, complete cds"
 M.musculus mRNA for NADP transhydrogenase
 "Mus musculus aldehyde dehydrogenase (ALDH2) mRNA, nuclear gene encoding mitochondrial protein
 "Mouse cytosolic epoxide hydrolase mRNA, complete cds"
 "Mus musculus 129SV carnitine palmitoyltransferase II mRNA, complete cds"
 "Mus musculus medium-chain acyl-CoA dehydrogenase mRNA, complete cds"
 "Mus musculus long-chain acyl-CoA dehydrogenase mRNA, complete cds"
 "Mus musculus very-long chain acyl-CoA dehydrogenase mRNA, partial cds"
 Mouse muscle creatine kinase mRNA (EC 2.7.3.2)
 "Mus musculus isocitrate dehydrogenase mRNA, complete cds"
 "Mus musculus long chain fatty acyl CoA synthetase mRNA, complete cds"
 "Mus musculus sterol carrier protein-2 (SCP-2) gene, complete cds"
 "Mouse alpha-tubulin isotype M-alpha-4 mRNA, complete cds"

Cardiomyopathy: Downregulated Genes

Anecdote vs. significance?



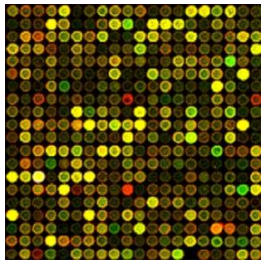
BLAST Definitions

"Homologous to sn 007021: PRE-MRNA SPLICING FACTOR SF2, P32 SUBUNIT PRECURSOR (GCIQ-R PROTEIN)
 "Mus musculus proteasome activator PA28 alpha subunit mRNA, complete cds"
 "Mus musculus cdc37 homolog mRNA, complete cds"
 "Mus musculus ornithine decarboxylase antisense gene, complete cds"
 M.musculus mRNA for carnitine acetyltransferase
 "Homologous to sn 000779: CALCIUM-TRANSPORTING ATPASE SARCOPLASMIC RETICULUM TYPE (EC 3.6.1.34)
 "Homologous to sn P11507: CALCIUM-TRANSPORTING ATPASE ENDOPLASTIC RETICULUM TYPE (EC 3.6.1.38)
 M.musculus ENO3 mRNA for enolase beta subunit
 "Homologous to sn P47858: 6-PHOSPHOFRUCTOKINASE, MUSCLE TYPE (EC 2.7.1.11) (PHOSPHOFRUCTOKINASE)
 Homologous to sn P23327: SARCOPLASMIC RETICULUM HISTIDINE-RICH CALCIUM-BINDING PROTEIN PRECURSOR
 "Mouse AE3 mRNA, complete cds"
 "M.musculus glucose transporter 2 mRNA, complete cds"
 Mus musculus aspartate aminotransferase gene 5'-flank and exon 1
 "Mus musculus thioredoxin-dependent peroxide reductase (txr) mRNA, complete cds"
 "Homologous to sn P47858: 6-PHOSPHOFRUCTOKINASE, MUSCLE TYPE (EC 2.7.1.11) (PHOSPHOFRUCTOKINASE)
 "Homologous to sn P11508: CALCIUM-TRANSPORTING ATPASE SARCOPLASMIC RETICULUM TYPE (EC 3.6.1.34)
 "Homologous to sn P35434: ATP SYNTHASE DELTA CHAIN, MITOCHONDRIAL PRECURSOR (EC 3.6.1.34)."
 "Mus musculus F1FO1 ATP synthase complex E subunit (Atp5k) gene, complete cds"
 "Mus musculus NAD(H)-specific isocitrate dehydrogenase gamma subunit precursor, mRNA, complete cds"
 "M.musculus gene for dodecenoyl-CoA delta-isomerase, exons 1 and 2"
 "Mus musculus cytochrome c oxidase subunit VIII-H precursor (COX8H) mRNA, complete cds"
 "Homologous to sn P35745: ACYLPHOSPHATASE, MUSCLE TYPE ISOZYME (EC 3.6.1.7) (ACYLPHOSPHATE PHOSPHATASE)
 "Mus musculus CD-1 cardiac troponin I mRNA, complete cds"
 "Homologous to sn P00566: CREATINE KINASE, M CHAIN (EC 2.7.3.2) (MU-2 PROTEIN)."
 Mouse mRNA for protein with homology to transition protein 2 (TP2)
 Mus musculus Selenium-binding liver protein mRNA
 "Mus musculus (clone MAR1) aldose reductase mRNA, complete cds"
 "Mus musculus vascular endothelial growth factor B 186 (VEGF-B) precursor, mRNA, complete cds"
 M.musculus mRNA for NADP transhydrogenase
 "Mus musculus aldehyde dehydrogenase (ALDH2) mRNA, nuclear gene encoding mitochondrial protein"
 "Mouse cytosolic epoxide hydrolase mRNA, complete cds"
 "Mus musculus 129SV carnitine palmitoyltransferase II mRNA, complete cds"
 "Mus musculus medium-chain acyl-CoA dehydrogenase mRNA, complete cds"
 "Mus musculus long-chain acyl-CoA dehydrogenase mRNA, complete cds"
 "Mus musculus very-long chain acyl-CoA dehydrogenase, partial cds"
 Mouse muscle creatine kinase mRNA (EC 2.7.3.2)
 "Mus musculus isocitrate dehydrogenase mRNA, complete cds"
 "Mus musculus long chain fatty acyl CoA synthetase mRNA, complete cds"
 "Mus musculus sterol carrier protein-2 (SCP-2) gene, complete cds"
 "Mouse alpha-tubulin isotype M-alpha-4 mRNA, complete cds"

Fatty Acid Degradation?
 Other pathways / processes?

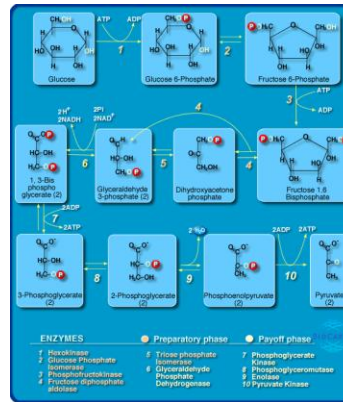
Interpreting Gene Lists

- My cool new screen worked and produced 1000 hits! ...Now what?
- Genome-Scale Analysis (Omics)
 - Genomics, Proteomics



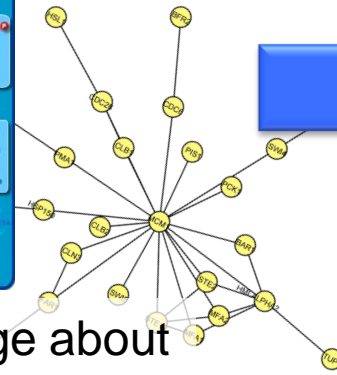
Ranking or clustering

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2



Prior knowledge about cellular processes

Analysis tools



Eureka! New heart disease gene!

Where Do Gene Lists Come From?

- Molecular profiling e.g. mRNA, protein
 - Identification → Gene list
 - Quantification → Gene list + values
 - Ranking, Clustering (biostatistics)
- Interactions: Protein interactions, Transcription factor binding sites (ChIP)
- Genetic screen e.g. of knock out library
- Association studies (Genome-wide)
 - Single nucleotide polymorphisms (SNPs)
 - Copy number variants (CNVs)

Other
examples?

What Do Gene Lists Mean?

- Biological system: complex, pathway
- Similar gene function e.g. protein kinase
- Similar cell or tissue location
- Chromosomal location (linkage, CNVs)

Biological Questions

- Step 1: What do you want to accomplish with your list (hopefully part of experiment design! 😊)
 - Summarize biological processes or other aspects of gene function
 - Controller for a process (TF)
 - Find new pathways or new pathway members
 - Discover new gene function
 - Correlation to a disease or phenotype (candidate gene prioritization)
 - Differential analysis – what's different between samples?

Biological Answers

- Computational analysis methods
 - Gene set analysis: summarize
 - Gene regulation network analysis
 - Pathway and network analysis
 - Gene function prediction
- But first! Gene list basics...

Gene Lists Overview

- Interpreting gene lists
- Gene attributes
 - Gene Ontology
 - Ontology Structure
 - Annotation
 - BioMart + other sources
- Gene identifiers and mapping

Gene Attributes

- Available in databases
- Function annotation
 - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
 - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
 - Splicing, 3' UTR, microRNA binding sites
- Protein properties
 - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Gene Attributes

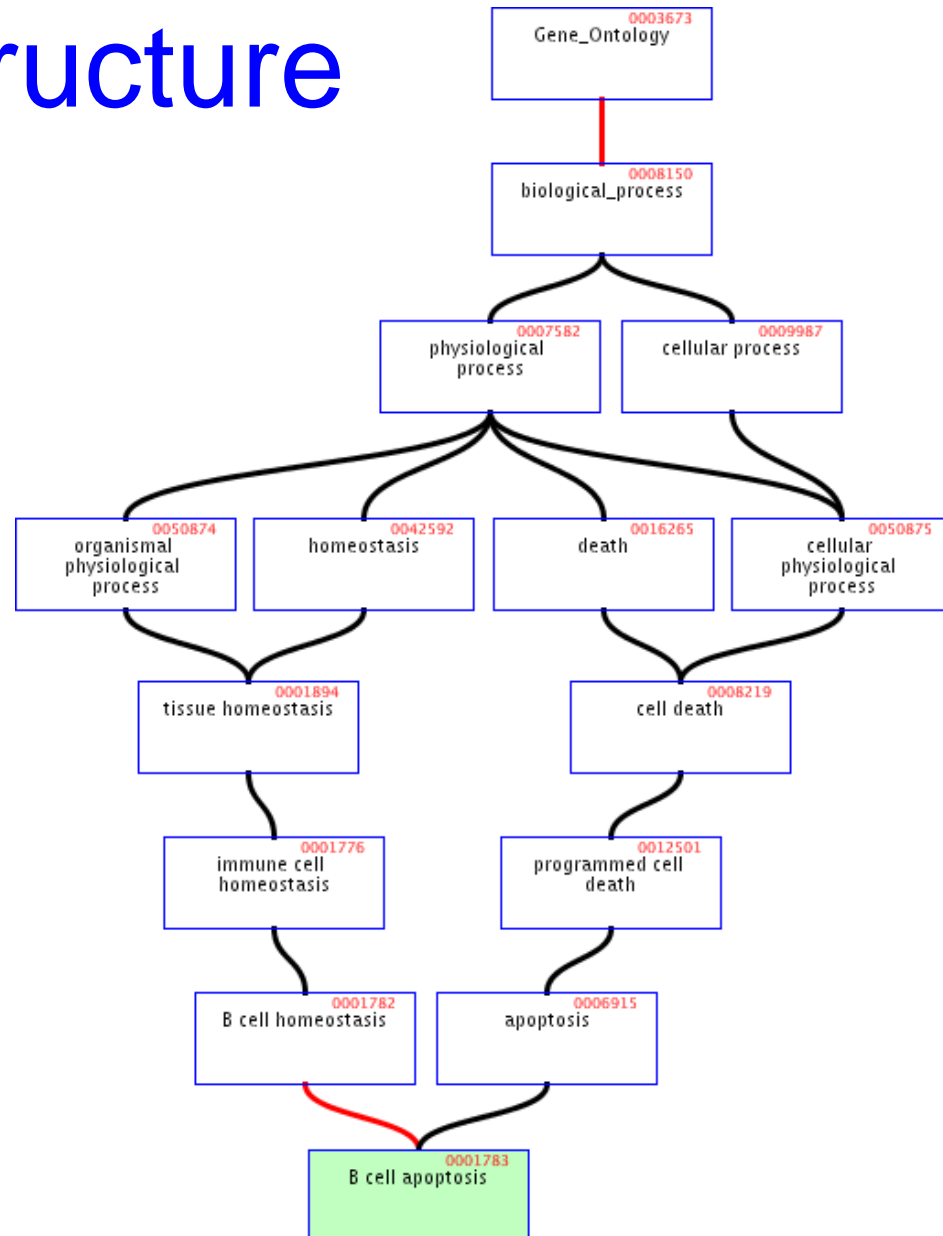
- Available in databases
- **Function annotation**
 - **Biological process, molecular function, cell location**
- Chromosome position
- Disease association
- DNA properties
 - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
 - Splicing, 3' UTR, microRNA binding sites
- Protein properties
 - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

What is the Gene Ontology (GO)?

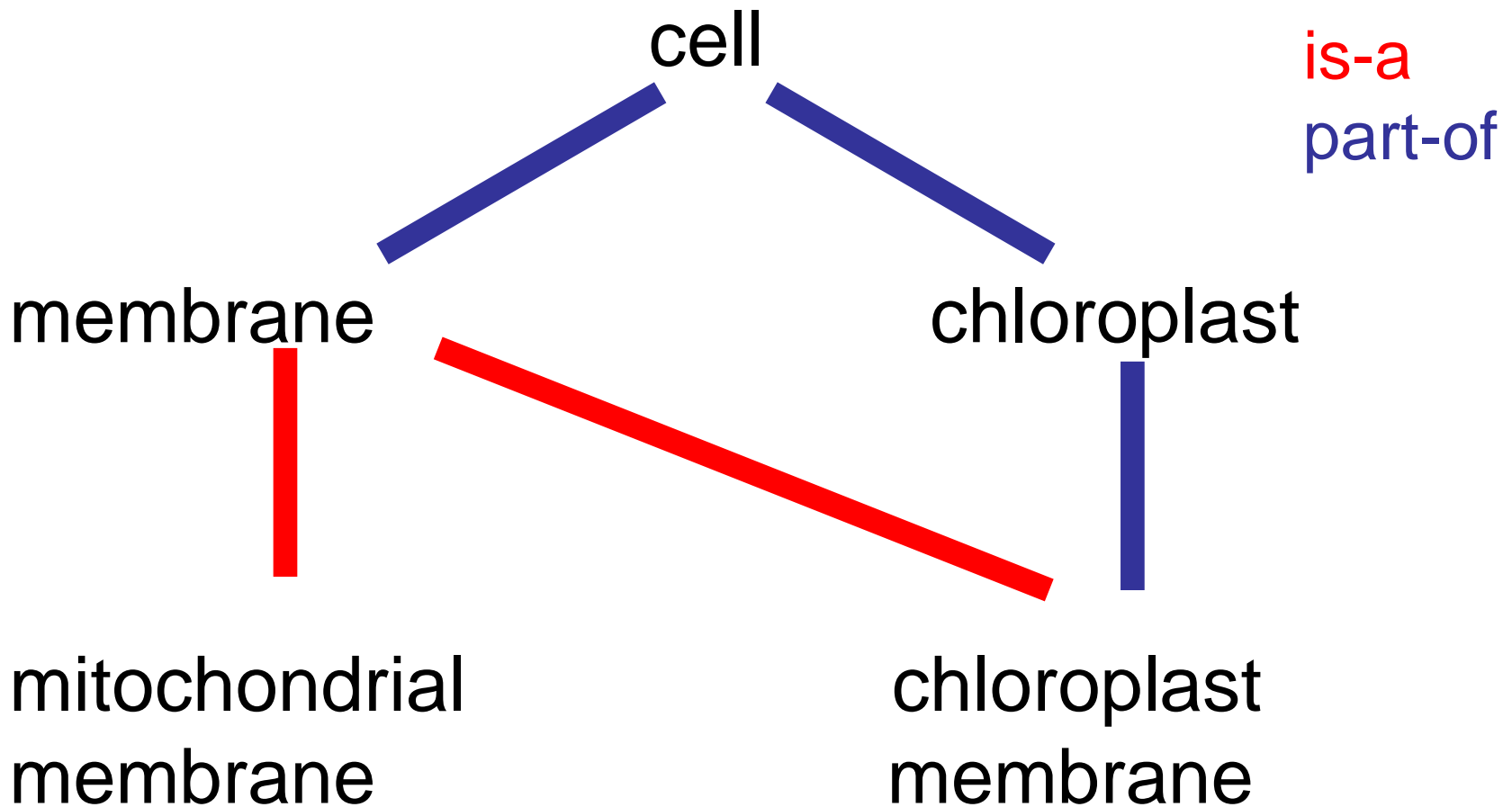
- Set of biological phrases (terms) which are applied to genes:
 - protein kinase
 - apoptosis
 - membrane
- Ontology: A formal system for describing knowledge

GO Structure

- Terms are related within a hierarchy
 - is-a
 - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



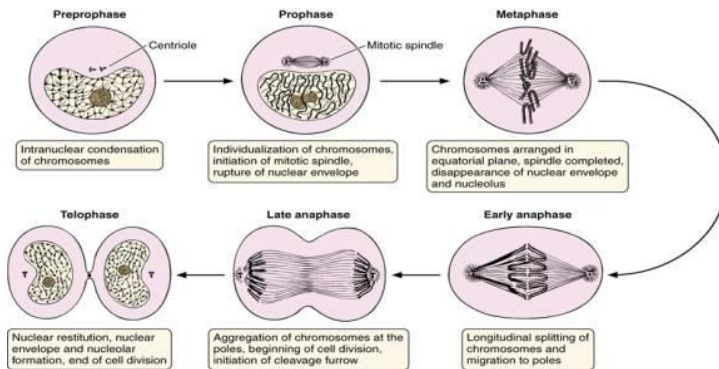
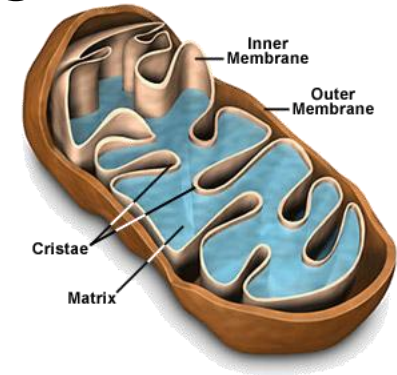
GO Structure



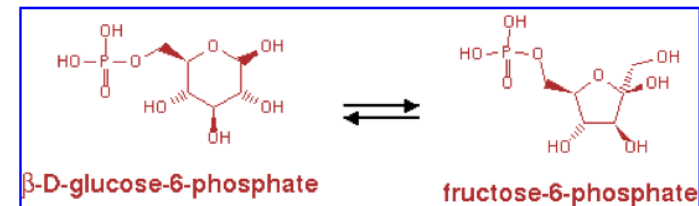
Species independent. Some lower-level terms are specific to a group, but higher level terms are not

What GO Covers?

- GO terms divided into three aspects:
 - cellular component
 - molecular function
 - biological process



Cell division



glucose-6-phosphate
isomerase activity

Terms

- Where do GO terms come from?
 - GO terms are added by editors at EBI and gene annotation database groups
 - Terms added by request
 - Experts help with major development
 - 27734 terms, 98.9% with definitions.
 - 16731 biological_process
 - 2385 cellular_component
 - 8618 molecular_function
 - As of July 6, 2009

Annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
 - Known as ‘gene associations’ or GO annotations
 - Multiple annotations per gene
- Some GO annotations created automatically

Annotation Sources

- Manual annotation
 - Created by scientific curators
 - High quality
 - Small number (time-consuming to create)
- Electronic annotation
 - Annotation derived without human validation
 - Computational predictions (accuracy varies)
 - Lower 'quality' than manual codes
- Key point: be aware of annotation origin

For your information

Evidence Types

- **ISS:** Inferred from Sequence/Structural Similarity
- **IDA:** Inferred from Direct Assay
- **IPI:** Inferred from Physical Interaction
- **IMP:** Inferred from Mutant Phenotype
- **IGI:** Inferred from Genetic Interaction
- **IEP:** Inferred from Expression Pattern
- **TAS:** Traceable Author Statement
- **NAS:** Non-traceable Author Statement
- **IC:** Inferred by Curator
- **ND:** No Data available



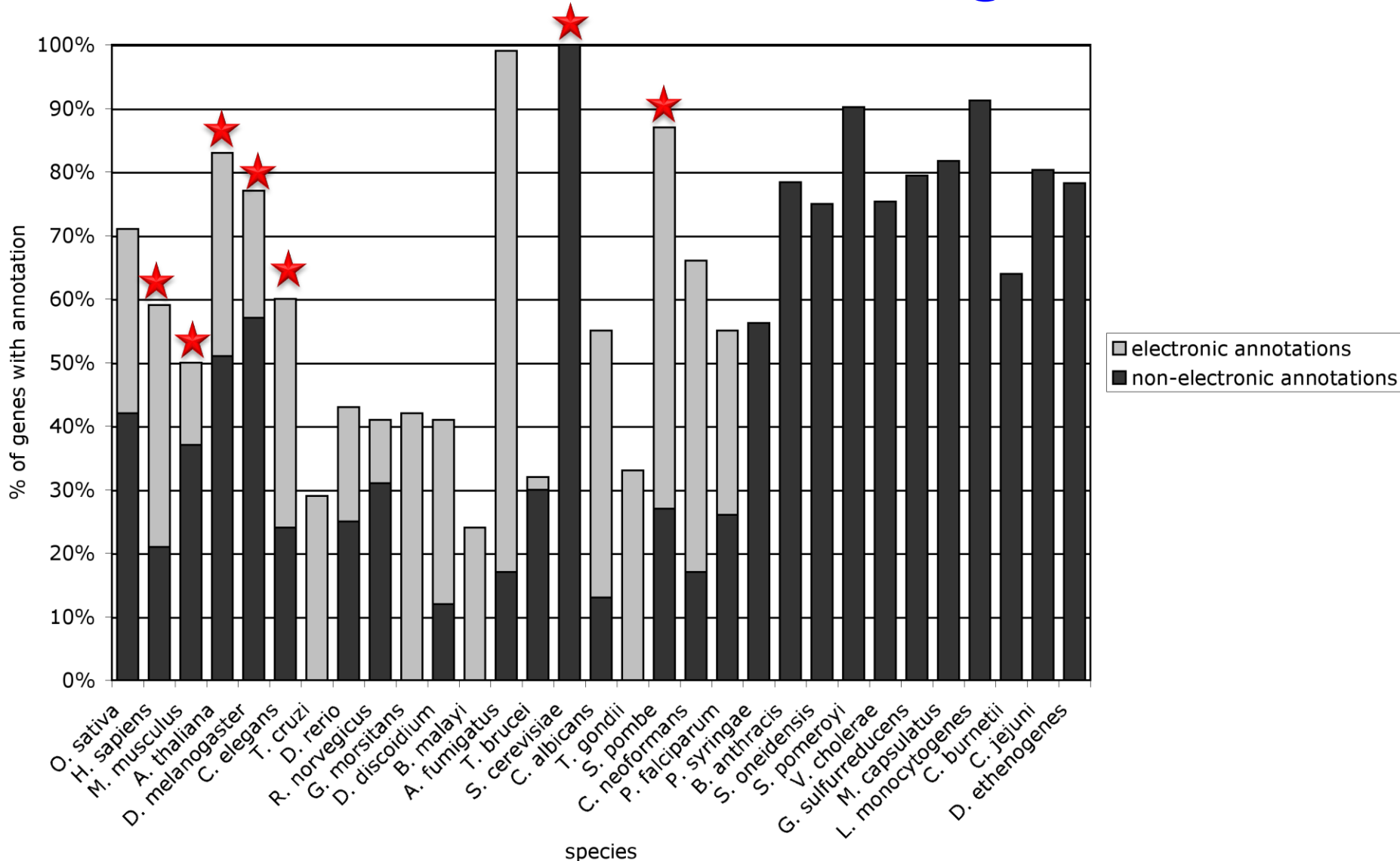
- **IEA:** Inferred from electronic annotation



Species Coverage

- All major eukaryotic model organism species
- Human via GOA group at UniProt
- Several bacterial and parasite species through TIGR and GeneDB at Sanger
- New species annotations in development

Variable Coverage



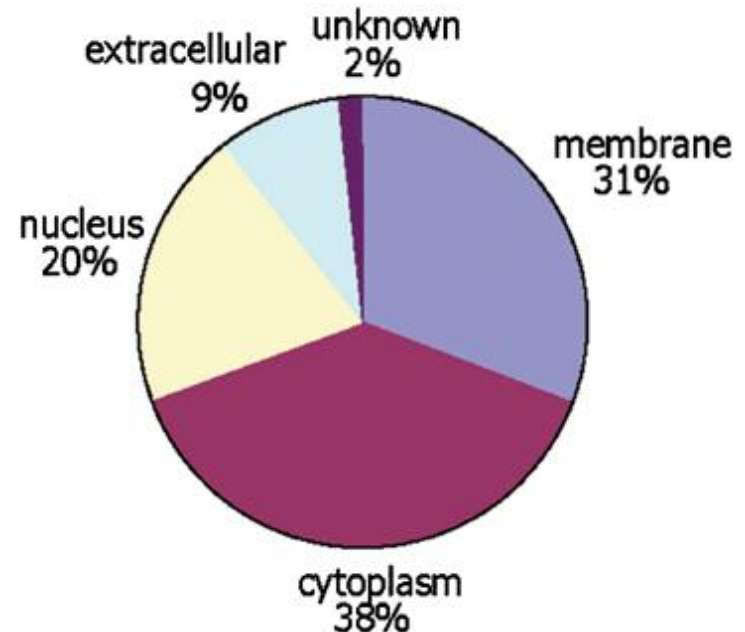
For your information

Contributing Databases

- [Berkeley *Drosophila* Genome Project \(BDGP\)](#)
- [dictyBase](#) (*Dictyostelium discoideum*)
- [FlyBase](#) (*Drosophila melanogaster*)
- [GeneDB](#) (*Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Leishmania major* and *Trypanosoma brucei*)
- [UniProt Knowledgebase](#) (Swiss-Prot/TrEMBL/PIR-PSD) and [InterPro](#) databases
- [Gramene](#) (grains, including rice, *Oryza*)
- [Mouse Genome Database \(MGD\) and Gene Expression Database \(GXD\)](#) (*Mus musculus*)
- [Rat Genome Database \(RGD\)](#) (*Rattus norvegicus*)
- [Reactome](#)
- [Saccharomyces Genome Database \(SGD\)](#) (*Saccharomyces cerevisiae*)
- [The Arabidopsis Information Resource \(TAIR\)](#) (*Arabidopsis thaliana*)
- [The Institute for Genomic Research \(TIGR\)](#): databases on several bacterial species
- [WormBase](#) (*Caenorhabditis elegans*)
- [Zebrafish Information Network \(ZFIN\)](#): (*Danio rerio*)

GO Slim Sets

- GO has too many terms for some uses
 - Summaries (e.g. Pie charts)
- GO Slim is an official reduced set of GO terms
 - Generic, plant, yeast



Crockett DK et al. Lab Invest. 2005
Nov;85(11):1405-15

GO Software Tools

- GO resources are freely available to anyone without restriction
 - Includes the ontologies, gene associations and tools developed by GO
- Other groups have used GO to create tools for many purposes
 - <http://www.geneontology.org/GO.tools>

Accessing GO: QuickGO

Search for a GO term: > examples - [apoptosis](#), [GO:0006915](#)

Search for a Protein: > examples - [tropomyosin](#), [P06727](#)

Compare GO terms: > example - [GO:0000122](#), [GO:0000001](#)

Find, view and download [annotation](#)

GO:0006915 apoptosis

A form of programmed cell death induced by external or internal signals that trigger the activity of proteolytic caspases, whose actions disintegrate the cell internally with condensation and subsequent fragmentation of the cell nucleus (blebbing) while the plasma membrane remains intact. Other features include the exposure of phosphatidyl serine on the cell surface.

[Term Information](#)

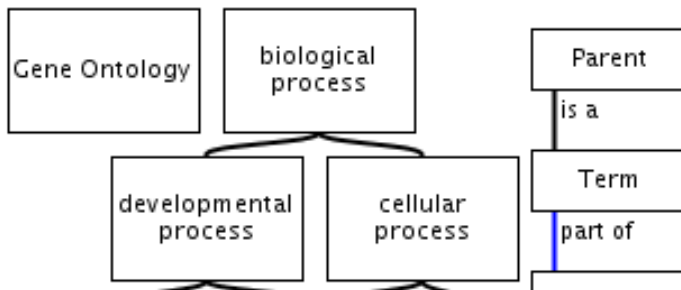
[Ancestor chart](#)

[Ancestor table](#)

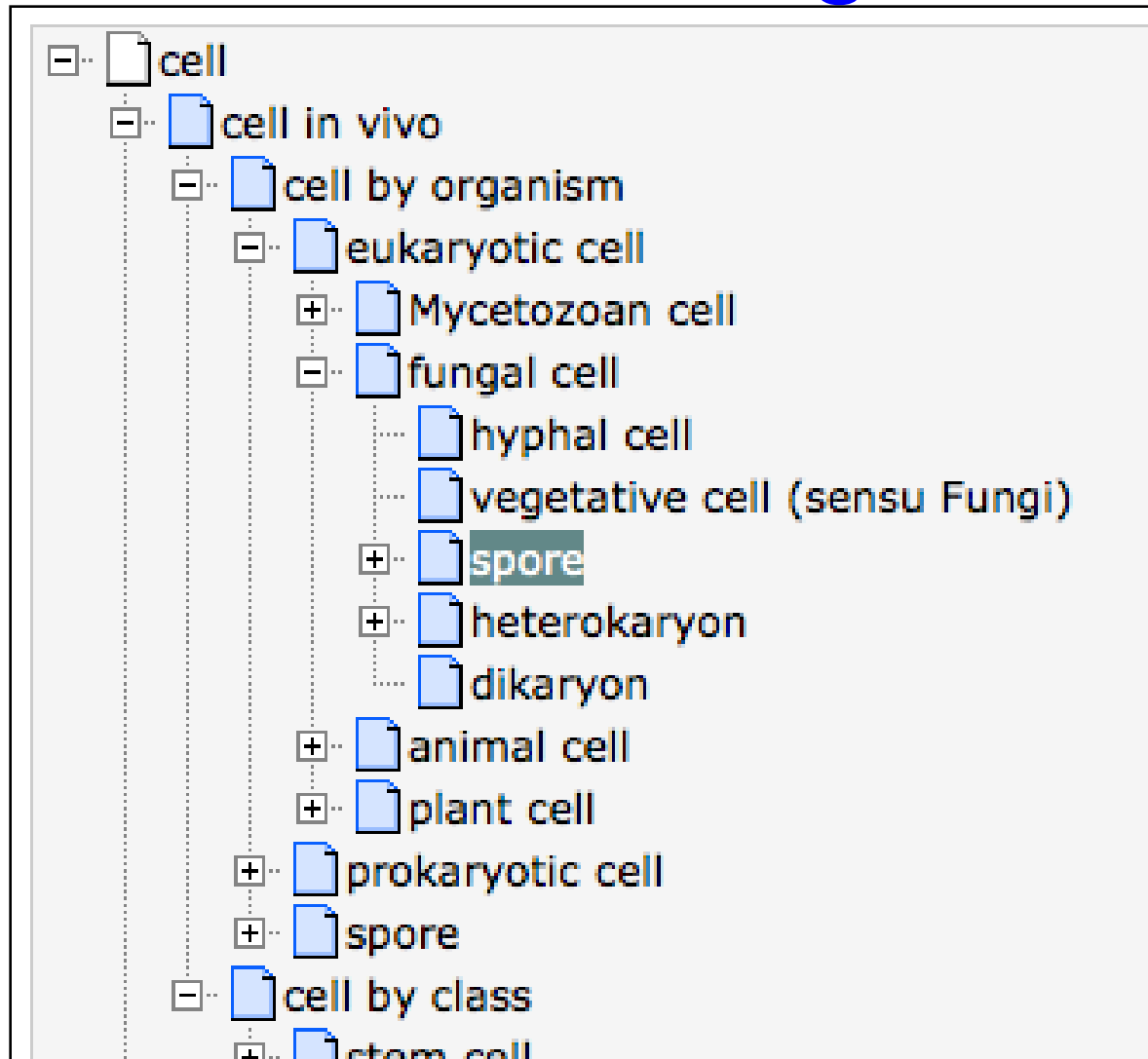
[Child Terms](#)

[Protein Annotation](#)

[Statistics](#)



Other Ontologies



Gene Attributes

- Function annotation
 - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
 - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
 - Splicing, 3' UTR, microRNA binding sites
- Protein properties
 - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Sources of Gene Attributes

- Ensembl BioMart (eukaryotes)
 - <http://www.ensembl.org>
- Entrez Gene (general)
 - <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>
- Model organism databases
 - E.g. SGD: <http://www.yeastgenome.org/>
- Many others: discuss during lab

Ensembl BioMart

- Convenient access to gene list annotation

The screenshot displays the Ensembl BioMart interface. On the left, a sidebar contains sections for 'Dataset', 'Filters', and 'Attributes'. The 'Dataset' section is highlighted in yellow and shows 'Ensembl Genes (release 49)'. The 'Filters' section shows '[None selected]'. The 'Attributes' section lists 'Ensembl Gene ID' and 'Ensembl Transcript ID'. The main area shows a dropdown menu for 'Homo sapiens genes (NCBI36)'. Below this, several filter categories are listed with checkboxes and options: 'REGION:', 'GENE:', 'GENE ONTOLOGY:', 'EXPRESSION:', 'MULTI SPECIES COMPARISONS:', 'PROTEIN:', and 'SNP:'. The 'SNP:' category includes sub-options like 'SNP IDs', 'Genes with SNPs that are', 'Synonymous status', and 'Associated with validated SNPs', each with a dropdown menu and radio buttons for 'Only' or 'Excluded'. On the right, three blue arrows point from the text 'Select genome', 'Select filters', and 'Select attributes to download' to the corresponding parts of the interface. At the bottom right, a section titled 'Features' and 'Homologs' lists various attributes for selection, including 'Structures', 'Sequences', 'SNPs', 'GENE:', 'EXTERNAL:', 'EXPRESSION:', 'PROTEIN:', 'GENOMIC REGION:', and 'Genomic Region Feature Attributes (clones etc.)' with sub-options like 'Feature chromosome', 'Feature chromosome start (bp)', 'Feature chromosome end (bp)', 'Feature class', 'Subtype category', and 'Subtype description'.

Dataset
Ensembl Genes (release 49)

Filters
[None selected]

Attributes
Ensembl Gene ID
Ensembl Transcript ID

Homo sapiens genes (NCBI36)

Select genome

Select filters

Select attributes to download

REGION:
 GENE:
 GENE ONTOLOGY:
 EXPRESSION:
 MULTI SPECIES COMPARISONS:
 PROTEIN:
 SNP:

SNP IDs SNPs with HGBASE ID(s) Only
 Excluded

Genes with SNPs that are Coding Only
 Excluded

Synonymous status Frameshifting SNPs Only
 Excluded

Associated with validated SNPs Only
 Excluded

Features Homologs
 Structures Sequences
 SNPs

GENE:
 EXTERNAL:
 EXPRESSION:
 PROTEIN:
 GENOMIC REGION:
Genomic Region Feature Attributes (clones etc.)
 Feature chromosome Feature class
 Feature chromosome start (bp) Subtype category
 Feature chromosome end (bp) Subtype description

What Have We Learned?

- Many gene attributes in databases
 - Gene Ontology (GO) provides gene function annotation
 - GO is a classification system and dictionary for biological concepts
 - Annotations are contributed by many groups
 - More than one annotation term allowed per gene
 - Some genomes are annotated more than others
 - Annotation comes from manual and electronic sources
 - GO can be simplified for certain uses (GO Slim)
- Many other gene attributes available from Ensembl and Entrez Gene

Gene Lists Overview

- Interpreting gene lists
- Gene function attributes
 - Gene Ontology
 - Ontology Structure
 - Annotation
 - BioMart + other sources
- Gene identifiers and mapping

Gene and Protein Identifiers

- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
 - E.g. Social Insurance Number, Entrez Gene ID 41232
- Gene and protein information stored in many databases
 - → Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
 - Important to recognize the correct record type
 - E.g. Entrez Gene records don't store sequence. They link to DNA regions, RNA transcripts and proteins.

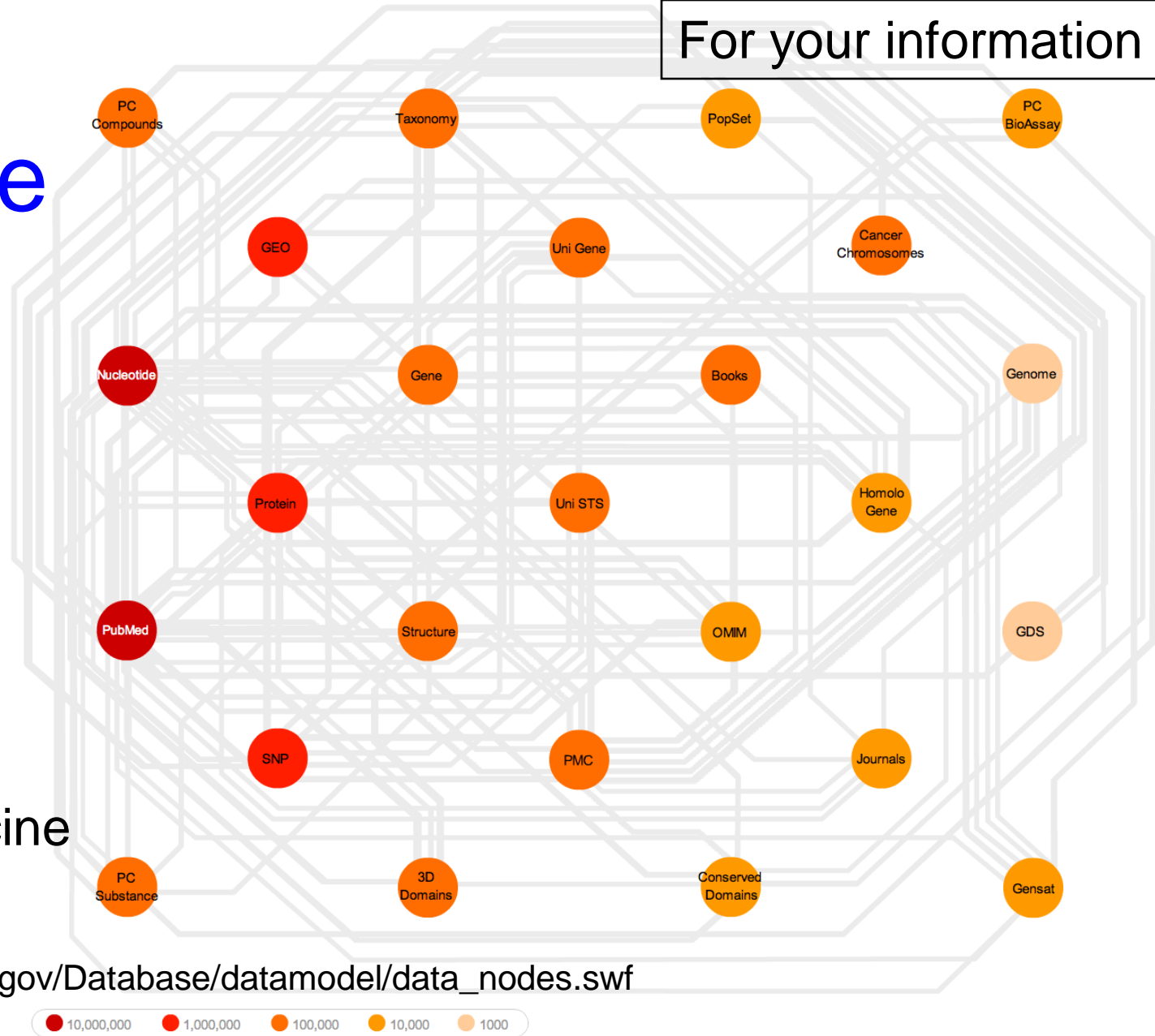
NCBI Database Links

NCBI:
U.S. National
Center for
Biotechnology
Information

Part of National
Library of Medicine
(NLM)

http://www.ncbi.nlm.nih.gov/Database/datamodel/data_nodes.swf

For your information



For your information

Common Identifiers

Gene

[Ensembl](#) ENSG00000139618

[Entrez Gene](#) 675

Unigene Hs.34012

RNA transcript

GenBank BC026160.1

[RefSeq](#) NM_000059

Ensembl ENST00000380152

Protein

Ensembl ENSP00000369497

[RefSeq](#) NP_000050.2

[UniProt](#) BRCA2_HUMAN or

A1YBP1_HUMAN

IPI IPI00412408.1

EMBL AF309413

PDB 1MIU

Species-specific

HUGO HGNC BRCA2

MGI MGI:109337

RGD 2219

ZFIN ZDB-GENE-060510-3

FlyBase CG9097

WormBase WBGene00002299 or ZK1067.1

SGD S000002187 or YDL029W

Annotations

InterPro IPR015252

OMIM 600185

Pfam PF09104

Gene Ontology GO:0000724

SNPs rs28897757

Experimental Platform

Affymetrix 208368_3p_s_at

Agilent A_23_P99452

CodeLink GE60169

Illumina GI_4502450-S

Red = Recommended

Identifier Mapping

- So many IDs!
 - Mapping (conversion) is a headache
- Four main uses
 - Searching for a favorite gene name
 - Link to related resources
 - Identifier translation
 - E.g. Genes to proteins, Entrez Gene to Affy
 - Unification during dataset merging
 - Equivalent records

ID Mapping Services

THE SYNERGIZER

The Synergizer database is a growing repository of gene and protein identifier synonym relationships. This tool facilitates the conversion of identifiers from one naming scheme (a.k.a "namespace") to another.

load sample inputs

Select species:

Select authority:

Select "FROM" namespace:

Select "TO" namespace:

(NB: The strings in [brackets] are representative IDs in the corresponding namespaces.)

File containing IDs to translate:

and/or

IDs to translate:

Output as spreadsheet:



*	entrezgene
YIL062C	854748
YLR370C	851085
YKL013C	853856
YNR035C	855771
YBR234C	852536

- Synergizer
 - <http://llama.med.harvard.edu/synergizer/translate/>
- Ensembl BioMart
 - <http://www.ensembl.org>
- UniProt
 - <http://www.uniprot.org/>

UniProt ID Mapping Service

UniProt - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.uniprot.org/

Most Visited Getting Started Latest Headlines

Stumble! I like it! All Share Info Favorites Friends Tools

Bioinformatics course - Donaldson ... UniProt Morris Lab at the University of Toro...

UniProt Downloads · Contact · Documentation/Help

Search in **Query**

Protein Knowledgebase (UniProtKB) Search Clear Fields

Search Blast Align Retrieve ID Mapping

WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none">★ Swiss-Prot, which is manually annotated and reviewed.★ TrEMBL, which is automatically annotated and is not reviewed.
UniRef	Sequence clusters, used to speed up similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations , taxonomy , keywords and more .

NEWS

UniProt release 15.7 – Sep 1, 2009
Formyl peptide receptors: the missing link between olfaction and immune system · Cross-references to STRING

- › Statistics for UniProtKB:
 - [Swiss-Prot](#) · [TrEMBL](#)
- › [Forthcoming changes](#)
- › [News archives](#)

SITE TOUR

Learn how to make best use of the tools and data on this site.

UniProt ID Mapping Service

The screenshot shows the UniProt website in a Mozilla Firefox browser. The address bar displays <http://www.uniprot.org/>. The browser's address bar also shows a search icon, a star icon, and the text "uniprot". The browser's menu bar includes File, Edit, View, History, Bookmarks, Tools, and Help. The browser's toolbar includes icons for Most Visited, Getting Started, Latest Headlines, Stumble!, I like it!, All, Share, Info, Favorites, Friends, and Tools. The browser's tabs include "Bioinformatics course - Donaldson ..." and "UniProt".

The UniProt website interface features a navigation bar with the UniProt logo and links for Downloads, Contact, and Documentation/Help. The main content area is divided into several sections:

- Identifiers:** A text input field containing "YIL062C".
- From:** A dropdown menu set to "UniProtKB AC/ID" with a "Map" button.
- To:** A dropdown menu set to "Entrez Gene (GeneID)" with a "Swap" button.
- or:** A text input field with a "Browse..." button.
- Clear:** A button to clear the input fields.
- Database identifier mapping tips:** A box containing the following text:

To map identifiers to or from UniProtKB:

 - enter identifiers, e.g.: 1TIA 1FNS
 - select a source database, e.g.: PDB
 - or select a target database, e.g.: UniProtKB

More...
- Search:** A button to search for the identifier.
- Blast:** A button to perform a BLAST search.
- Align:** A button to align sequences.
- Retrieve:** A button to retrieve sequence data.
- ID Mapping:** A button to perform ID mapping.

WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none">★ Swiss-Prot, which is manually annotated and reviewed.★ TrEMBL, which is automatically annotated and is
------------------	--

NEWS

UniProt release 15.7 – Sep 1, 2009

Formyl peptide receptors: the missing link between olfaction and immune system · Cross-references to STRING

- › Statistics for UniProtKB:
 - Swiss-Prot · TrEMBL
- › Forthcoming changes
- › News archives

SITE TOUR

ID Mapping Challenges

- Avoid errors: map IDs correctly
- Gene name ambiguity – not a good ID
 - e.g. FLJ92943, LFS1, TRP53, p53
 - Better to use the standard gene symbol: TP53
- Excel error-introduction
 - OCT4 is changed to October-4
- Problems reaching 100% coverage
 - E.g. due to version issues
 - Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80

ID Mapping Challenges

- Spot-test any ID mapping service you use.
- Check samples from first, last and middle of your list of identifiers to be converted.
- Ask for help if you are uncertain. This is a field of expertise in itself.
- A new plugin for Cytoscape has been released called CyThesaurus based on [bridgedb.org](http://www.bridgedb.org).
- Programmers see also, <http://www.bridgedb.org>

What Have We Learned?

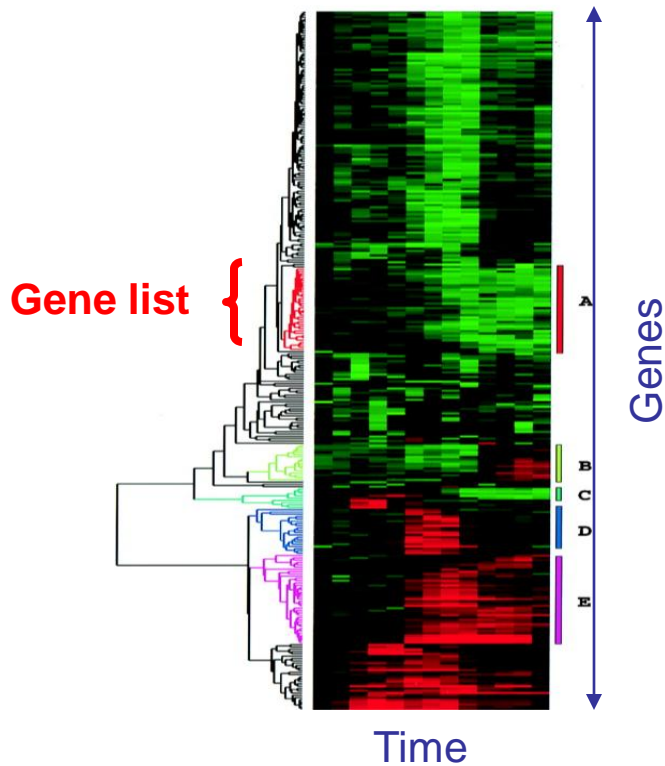
- Genes and their products and attributes have many identifiers (IDs)
- Genomics requirement to convert or map IDs from one type to another
- ID mapping services are available
- Use standard, commonly used IDs to avoid ID mapping challenges
- **Spot-test ID mapping services you use**

Break

- Over-representation analysis next

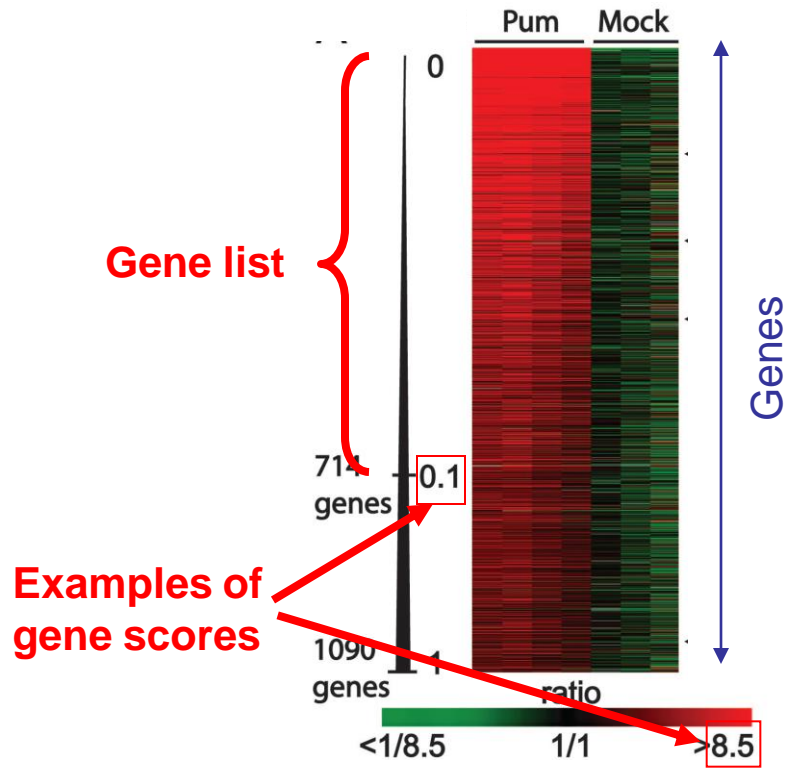
Examples of sources of gene lists

Clustering



Source Eisen et al. (1998) PNAS 95

Thresholding a gene "score"



Source: Gerber et al. (2006) PNAS103

Over-representation analysis (ORA) in a nutshell

- Given:
 1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast), or Gene Scores: RRP6 (4.0), MRD1 (3.0) etc
 2. Gene annotations: e.g. Gene ontology, transcription factor binding sites in promoter
- ORA Question: *Are any of the gene annotations surprisingly enriched in the gene list?*
- Details:
 - How to assess “surprisingly” (statistics)
 - How to correct for repeating the tests

Overview

- Theory:
 - Review: What is a P-value? The good ole' T-test.
 - Fisher's Exact Test, the bread and butter of ORA
 - Correcting for multiple testing
 - Enrichment analysis with gene rankings

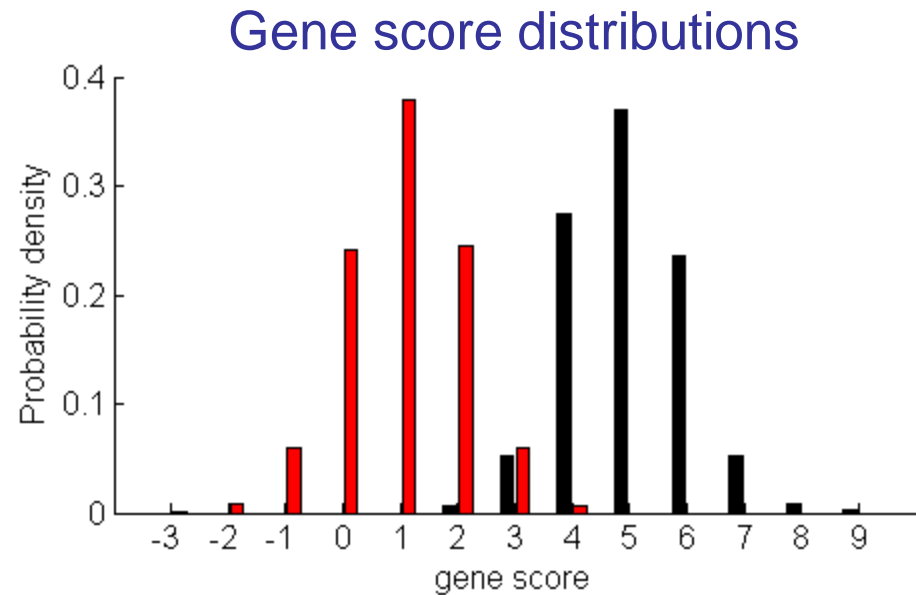
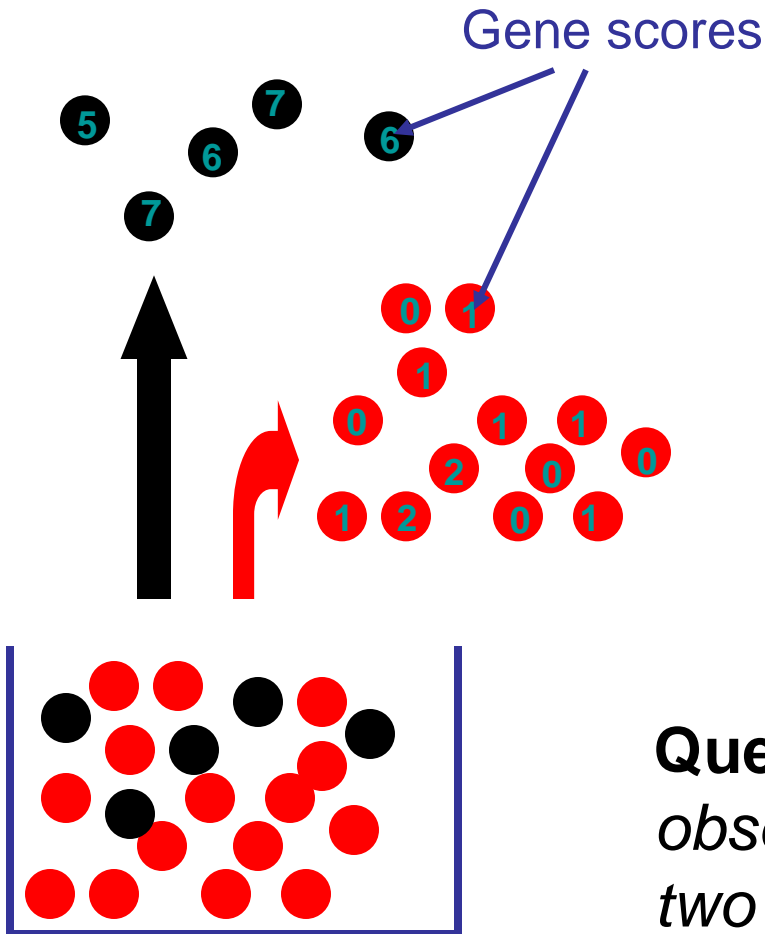
Overview

- Theory:
 - Review: What is a P-value? The good ole' T-test.
 - Fisher's Exact Test, the bread and butter of ORA
 - Correcting for multiple testing
 - Enrichment analysis with gene rankings

What is a P-value?

- The P-value is (a bound) on the probability that the “null hypothesis” is true,
- Calculated by calculating **statistics** using the data and testing the probability of observing those statistics, or ones more extreme, given a sample of the same size distributed according to the null hypothesis,
- Intuitively: *P-value is the probability of a false positive result* (aka “Type I error”)

The good old T-test



Question: *How likely are the observed differences between the two distributions due to chance?*

ORA using the T-test

Answer: *Two-tailed T-test*

Black: $N_1=500$

Mean: $m_1 = 1.1$

Std: $s_1 = 0.9$

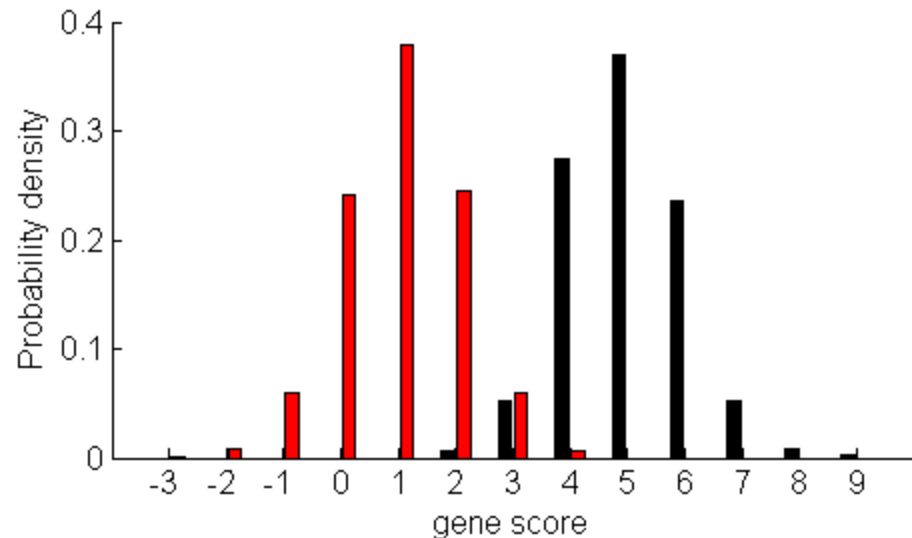
Red: $N_2=4500$

Mean: $m_2 = 4.9$

Std: $s_2 = 1.0$

$$\begin{aligned} \text{T-statistic} &= \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \\ &= -88.5 \end{aligned}$$

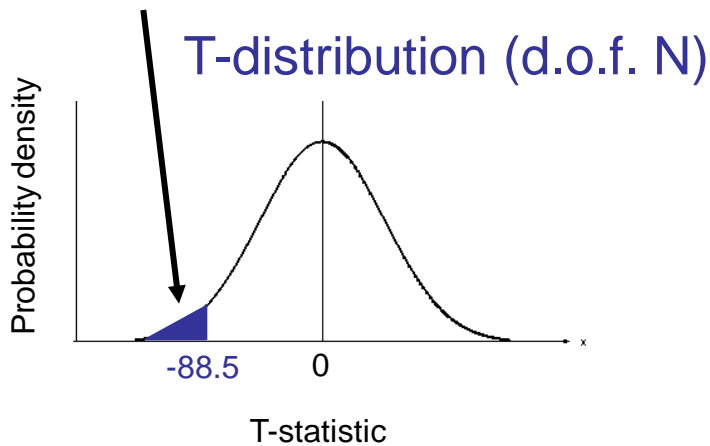
Gene score distributions



Formal Question: *What is the probability of observing the T-statistic or one more extreme if the means of the two distributions were the same?*

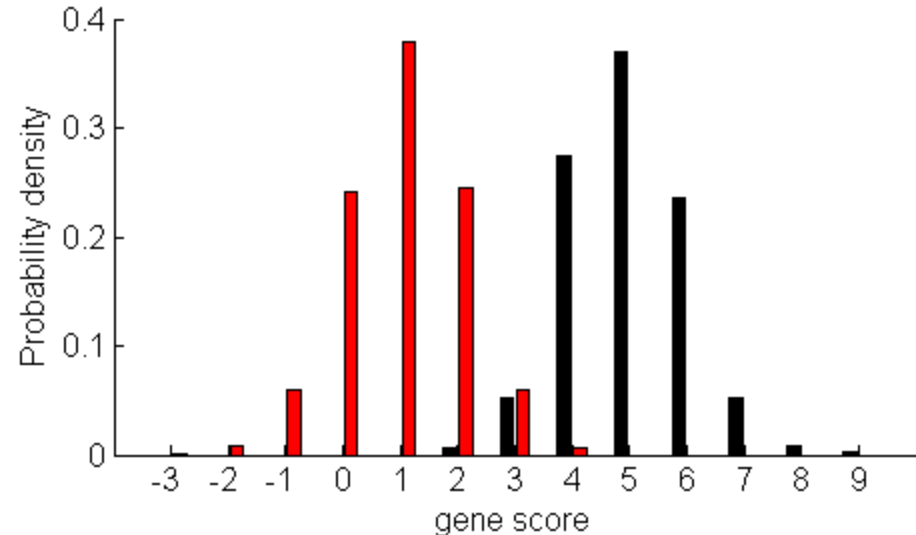
ORA using the T-test

P-value = shaded area * 2



$$\begin{aligned} \text{T-statistic} &= \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \\ &= -88.5 \end{aligned}$$

Gene score distributions



Formal Question: *What is the probability of observing the T-statistic or one more extreme if the means of the two distributions were the same?*

Overview

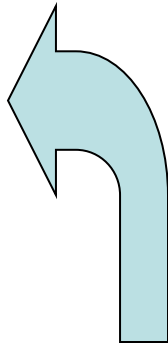
- Theory:
 - Review: What is a P-value? The good ole' T-test.
 - Fisher's Exact Test, the bread and butter of ORA
 - Correcting for multiple testing
 - Enrichment analysis with gene rankings

Fisher's exact test: the bread and butter of ORA

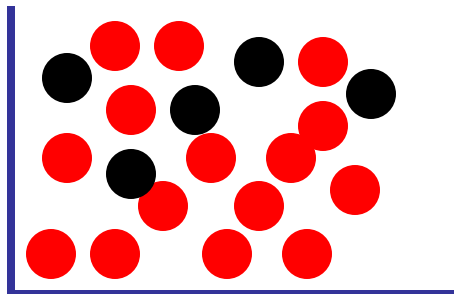
a.k.a., the hypergeometric test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



Formal question: *What is the probability of finding 4 or more black genes in a random sample of 5 genes?*

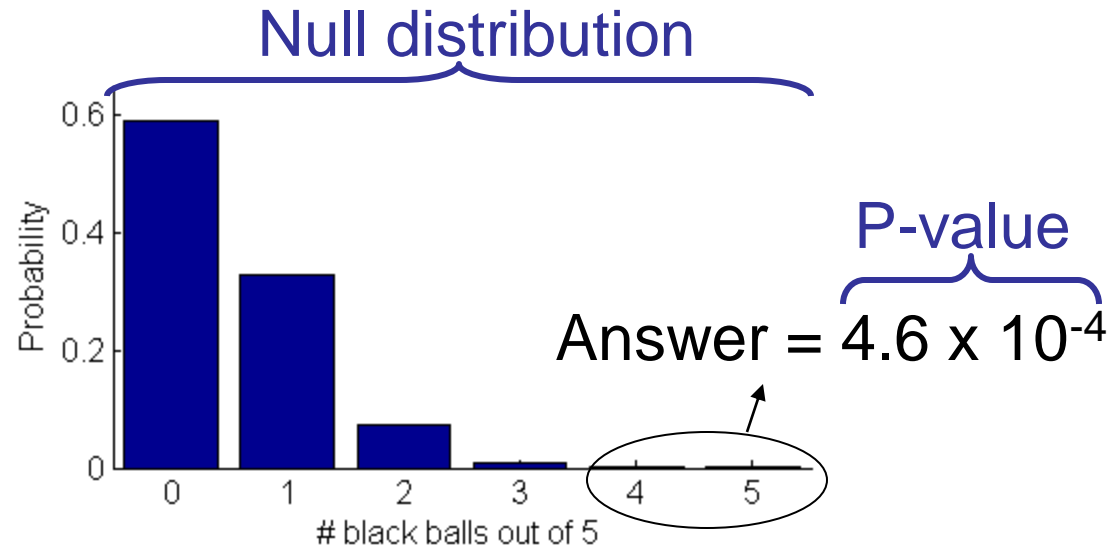
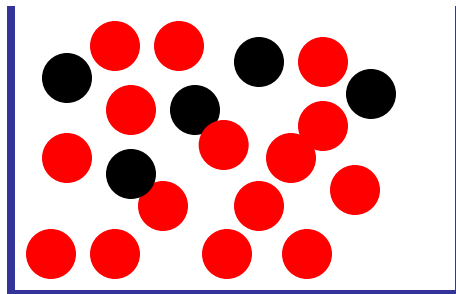
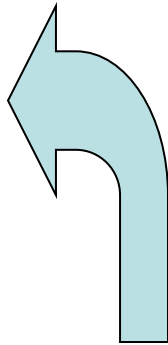


Background population:
500 black genes,
5000 red genes

Fisher's exact test cont.

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



Background population:
500 black genes,
5000 red genes

Important details

- To test for *under-enrichment* of “black”, test for *over-enrichment* of “red”.
- Need to choose “background population” appropriately, e.g., if only portion of the total gene complement is queried (or available for annotation), only use that population as background.
- To test for enrichment of more than one independent types of annotation (red vs black and circle vs square), apply Fisher’s exact test separately for each type. ***More on this later***

What have we learned?

- Fisher's exact test is used for ORA of gene lists for a single type of annotation,
- P-value for Fisher's exact test
 - is “the probability that a random draw of the same size as the gene list from the background population would produce the observed number of annotations in the gene list or more.”,
 - and depends on size of both gene list and background population as well and # of black genes in gene list and background.

Overview

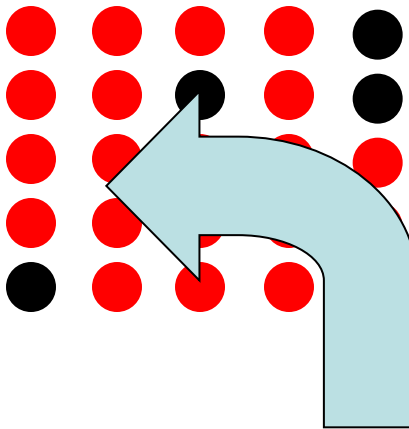
- Theory:
 - Review: What is a P-value? The good ole' T-test.
 - Fisher's Exact Test, the bread and butter of ORA
 - Correcting for multiple testing
 - Enrichment analysis with gene rankings

Correcting for multiple testing: overview

- Why do we need to correct? Winning the P-value lottery.
- Controlling the Family-wise Error Rate (FWER) with the Bonferroni-correction
- Controlling the false-discovery rate (FDR): Benjamini-Hochberg, Storey-Tibshirani, Q-values and all that

How to win the P-value lottery, part 1

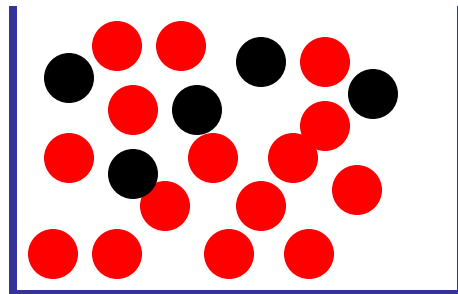
Random draws



... 7,834 draws later ...



*Expect a random draw
with observed
enrichment once every
 $1 / P\text{-value}$ draws*



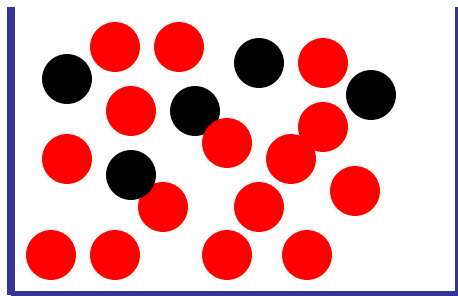
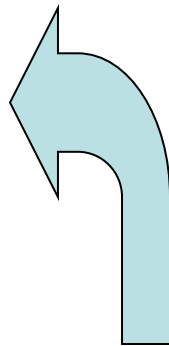
Background population:
500 black genes,
5000 red genes

How to win the P-value lottery, part 2

Keep the gene list the same, evaluate different annotations

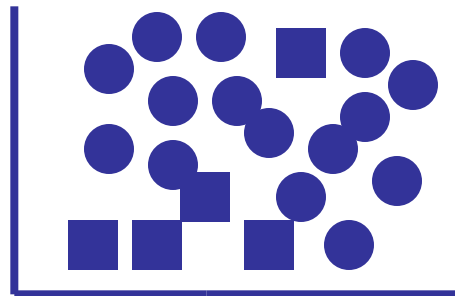
Observed draw

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



Different annotations

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



ORA tests need correction

From the Gene Ontology website:

Current ontology statistics: **25206** terms

- **14825** biological process
- **2101** cellular component
- **8280** molecular function

Two types of multiple test corrections

- Controlling the **Family-Wise Error Rate (FWER)** controls the probability that any test is a false positive
- Controlling the **False Discovery Rate (FDR)** controls the proportion of positive tests (i.e. rejections of the null hypothesis) that are false positives

Controlling Family-Wise Error Rate using the Bonferroni correction

If $M = \#$ of annotations tested:

Corrected P-value = $M \times$ original P-value

Corrected P-value is greater than or equal to the probability that any single one of the observed enrichments could be due to random draws. The jargon for this correction is “**controlling for the *Family-Wise Error Rate (FWER)***”

Bonferroni correction caveats

- Bonferroni correction is very stringent and can “wash away” real enrichments.
- Often users are willing to accept a less stringent condition, the “false discovery rate” (FDR), which leads to a gentler correction when there are real enrichments.

False discovery rate (FDR)

- FDR is *the expected **proportion** of the observed enrichments that are due to random chance.*
- Compare to Bonferroni correction which is *the probability that **any one** of the observed enrichments is due to random chance.*

Benjamini-Hochberg (B-H) FDR

If α is the desired FDR (ie level of significance), then choose the corresponding cutoff for the original P-values as follows:

1) Rank all “M” P-values

2) Test each P-value against

$$q = \alpha \times (M - \text{Rank} + 1) / M$$

e.g. Let $M = 100$, $\alpha = 0.05$

P-value	Rank	q	Is P-value < q?
0.9	1	0.05 X 1.00	No
0.7	2	0.05 x 0.99	No
0.5	3	0.05 X 0.98	No
0.04	4	0.05 x 0.97	Yes
...
0.005	M	0.05 x 0.01	No

3) New P-value cutoff, i.e. “ α ”, is lowest ranked P-value to pass the test.

P-value cutoff of 0.04 ensures FDR < 0.05

Reducing multiple test correction stringency

- The correction to the P-value threshold α depends on the # of tests that you do, so, no matter what, the more tests you do, the more sensitive the test needs to be
- Can control the stringency by reducing the number of tests: e.g. use GO slim or restrict testing to the appropriate GO annotations.

What have we learned

- When testing multiple annotations, need to correct the P-values (or, equivalently, α) to avoid winning the P-value lottery.
- There are two types of corrections:
 - **Bonferroni** controls the probability any one test is due to random chance (aka FWER) and is very stringent
 - **B-H** controls the FDR, i.e., expected proportion of “hits” that are due to random chance
- Can control stringency by carefully choosing which annotation categories to test.

DAVID, part 1

<http://david.abcc.ncifcrf.gov/>

Paste list here

DAVID automatically detects organism

Choose ID type

List type: list or background?

DAVID: Functional Annotation Tools - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://david.abcc.ncifcrf.gov/tools.jsp

Getting Started Latest Headlines

Analysis Wizard
DAVID Bioinformatics Resources 2008, NIAID/NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID?

Upload List Background

Upload Gene List

[Demolist 1](#) [Demolist 2](#)

[Upload Help](#)

Step 1: Enter Gene List

A: Paste a list

Clear

Or

B: Choose From a File

Browse...

Step 2: Select Identifier

AFFY_ID

Step 3: List Type

Gene List

Background

Step 1. Successfully submitted gene list
Current Gene List: Uploaded List_1
Current Background: SACCHAROMYCES CEREVISIAE

Step 2. Analyze above gene list with one of DAVID tools

Functional Annotation Tool

- Functional Annotation Clustering
- Functional Annotation Chart
- Functional Annotation Table

Gene Functional Classification Tool

Gene ID Conversion Tool

Gene Name Batch Viewer

Done

Slide 43 of 50 Bioinformatics Master

Start Microsoft PowerPoint - ... DAVID: Functional A... Help - Mozilla Firefox 100%

DAVID, part 2

<http://david.abcc.ncifcrf.gov/>

The screenshot shows the DAVID Functional Annotation Tool interface. The browser window title is "DAVID: Functional Annotation Result Summary - Mozilla Firefox". The address bar shows the URL "http://david.abcc.ncifcrf.gov/summary.jsp". The page header includes the DAVID logo and the text "Functional Annotation Tool" and "DAVID Bioinformatics Resources 2008, NIAID/NIH". A navigation menu contains links for Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, Why DAVID?, and About Us.

The main content area is titled "Annotation Summary Results" and includes a "Help and Tool Manual" link. It displays the following information:

- Current Gene List: Uploaded List_1
- 860 DAVID IDs
- Current Background: SACCHAROMYCES CEREVISIAE (checked) with a "Check Defaults" button and a "Clear All" button.

A list of selected annotations is shown with checkboxes:

- Main Accessions (0 selected)
- Other Accessions (0 selected)
- Gene Ontology (3 selected)
- Protein Domains (3 selected)
- Pathways (3 selected)
- General Annotations (0 selected)
- Functional Categories (3 selected)
- Protein Interactions (0 selected)
- Literature (0 selected)
- Disease (1 selected)
- Tissue Expression

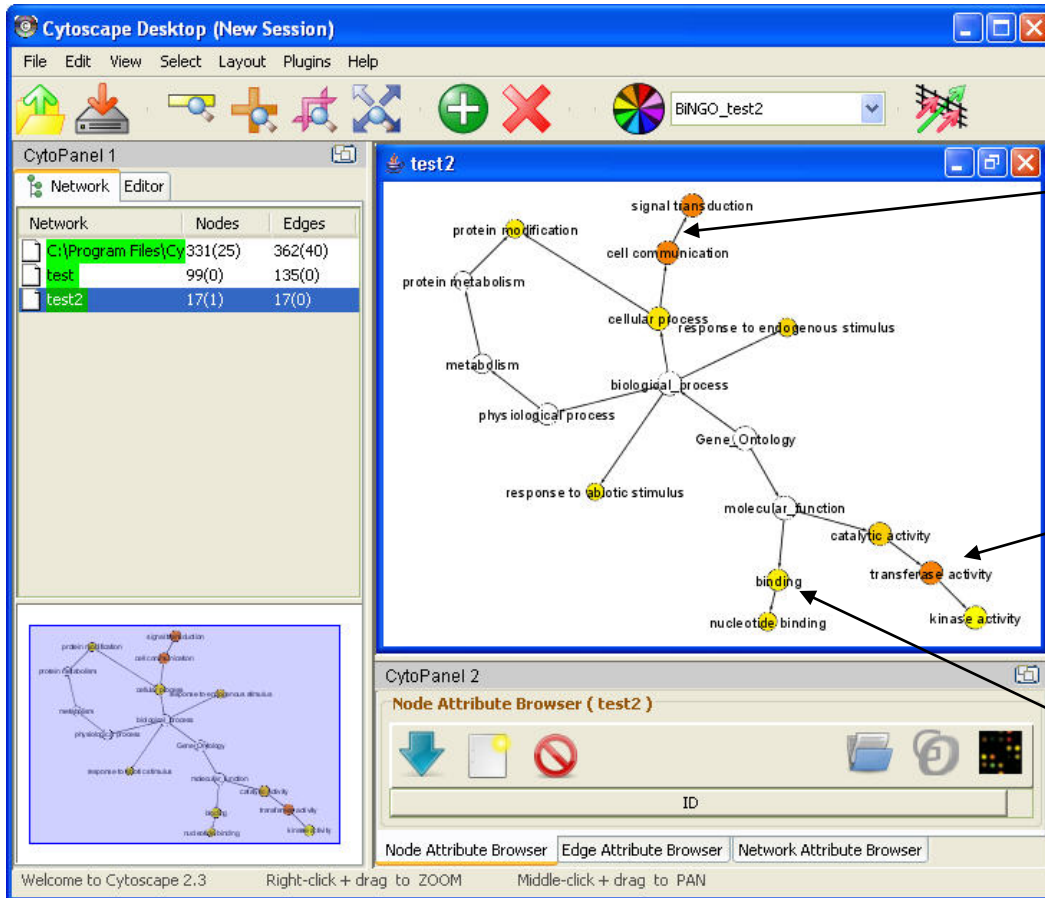
Below this list is a section titled "Combined View for Selected Annotation" with three buttons: "Functional Annotation Clustering^{new!}" (highlighted with a red arrow), "Functional Annotation Chart", and "Functional Annotation Table".

On the left side, there is a "Gene List Manager" section with a dropdown menu for species selection (currently set to "SACCHAROMYCES CEREVISIAE") and a "List Manager" section with a dropdown menu for list selection (currently set to "Uploaded List_1").

The Windows taskbar at the bottom shows the Start button, open applications (Microsoft PowerPoint, DAVID: Functional A..., Help - Mozilla Firefox), system tray icons, and the time "9:03 PM".

BINGO, an ORA cytoscape plugin

<http://www.psb.ugent.be/cbd/papers/BiNGO/index.htm>



Links represent parent-child relationships in GO ontology

Colours represent significance of enrichment

Nodes represent GO categories

GoMiner, part 1

<http://discover.nci.nih.gov/gominer>

GO MINER™ Genomics and Bioinformatics Group
LMP, CCR, National Cancer Institute

Home High-Throughput Getting Started Requirements Installation Downloads Command Line Database FAQ News Citings GoMiner Papers Credits

High-Throughput Home Web Interface Web Input Command Line Interface Command Line Input Output Files Examples Using CIMminer Error Codes Process Overview Supplementary Materials

This is the web interface for High-Throughput GoMiner™. You need to upload two files to the server, a total file with all of the genes in your experiment, and one of two types of changed files. Detailed descriptions of the [input files](#) are available. You will receive an email message with instructions on how to download your results once they are complete. Documentation for both the [output files](#) generated and possible [error codes](#) are also online.

Step 1: Select total file
Input should be a list file with ".txt" extension. This file is specified:

OR
Select [Auto-Generate](#) option
 [Auto-Generate](#) (Increases Computation Time Slightly)

Step 2: Select the changed file
Acceptable file with ".txt" or ".xls" file with ".txt" extension

Step 3: Select DataSource(s)
 Choose from list
UniProt (H. sapiens et al.)
TIGR_TGI (H. sapiens et al.)
TIGR_CMR (Microbes)
FB (D. melanogaster)

OR
 Specify semicolon separated list of data sources. You find possible values at the [Gene Ontology web site](#); look at the abbreviate and synonym fields.
For example, UniProt and Mouse Consensus would be UniProt;MOU

Step 4: Select Organism(s)
 Choose from list
H. sapiens
M. musculus
R. norvegicus
D. melanogaster

OR
 Specify semicolon separated [NCBI Tax IDs](#)
For example, Human and mouse would be given as: 9606;10090

Step 5: Select Evidence Code(s)
 Choose from list
Evidence Level 1 (TAS,IDA,IMP,IGI,JP,ISS,JEP,NAS,RCA)
Evidence Level 2 (TAS,IDA,IMP,IGI,JP,ISS,JEP,NAS)
Evidence Level 3 (TAS,IDA,IMP,IGI,JP,ISS,JEP)
Evidence Level 4 (TAS,IDA,IMP,IGI,JP)

OR
 Specify semicolon separated [GO Evidence Codes](#)
For example, "Inferred by Curator" and "Inferred from Direct Observation" would be given as: IC:JBA

Step 6: Select Lookup Settings
 Enhanced Names (UniProt Only)
Enhanced Names only affect results if UniProt or All Data Sources is selected in Step 3.
 Cross Reference
 Synonym

1. Click "web interface"

2. Upload names of background genes

3. Upload gene list

4. Choose organism

5. Choose evidence code (All or Level 1)

GoMiner, part 2

Step 7: Select Statistical Constraints for Summary Reports

P-Value
 FDR
 Both

The p-value and/or FDR determine which categories are counted in the report file and which categories are displayed in the GUI.

Step 8: Select number of randomizations

Step 9: Smallest Category Size for Category Statistics

Limits the categories that will be included in the CIM and summary reports. Categories whose size (i.e., the number of genes) is less than this threshold will be omitted from category statistics calculations, and randomized categories below this threshold will be omitted from FDR calculations.

Step 10: CIM

No CIM
 Basic CIM
 Additional CIMs (Uses Moderately Increased Hard Drive Storage)

Largest Category Size (i.e., the number of genes) to Include in CIM

Step 12: Select Root Category for evaluating enrichment ratios, Fisher's Exact and the FDR's.

All/Gene ontology
 Biological process
 Cellular component
 Molecular function

Step 12: [TF Binding](#) (currently supports only Human/Mouse and Rat Organism's)

Compute TF Binding Sites at [ABCC](#) (Increases Computation Time Significantly)

Step 13: Your E-mail Address

Step 13: Submit your query

We would like to hear from you. You can reach the team via [email](#).

GoMiner was originally developed jointly by the [Genomics and Bioinformatics Group \(GEG\)](#) of LMP, NCI, NIH and the [Medical Informatics and Bioinformatics group](#) of EME, Georgia Tech/Emory University. It is now maintained and under continuing development by GEG.

[Notice and Disclaimer](#)

Done

Slide 35 of 36 Bioinformatics Master English (U.S.)

6. Restrict # of tests via category size

7. Restrict # of tests via GO hierarchy

8. Results emailed to this address, in a few minutes

Break

- Try out an over-representation analysis with multiple test correction using DAVID.

Overview

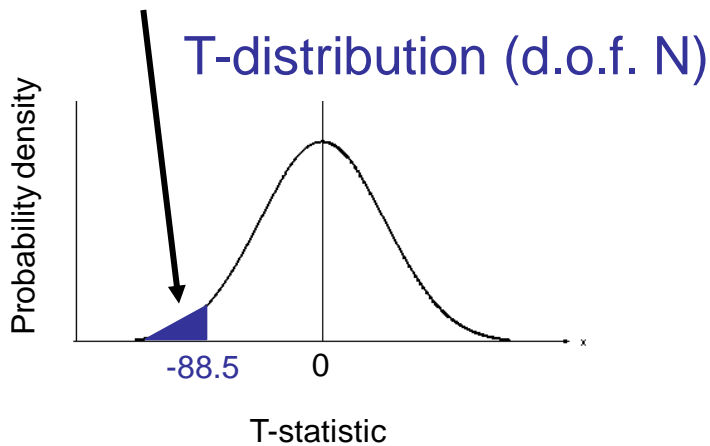
- Theory:
 - Review: What is a P-value? The good ole' T-test.
 - Fisher's Exact Test, the bread and butter of ORA
 - Correcting for multiple testing
 - Enrichment analysis with gene rankings

ORA with gene rankings: overview

- Why can't I use the T-test?
- The Wilcoxon-Mann-Whitney (WMW) test: a T-test on ranks
- The Kolmogorov-Smirnov (KS) test: testing for arbitrary differences between gene score distributions.
- GSEA

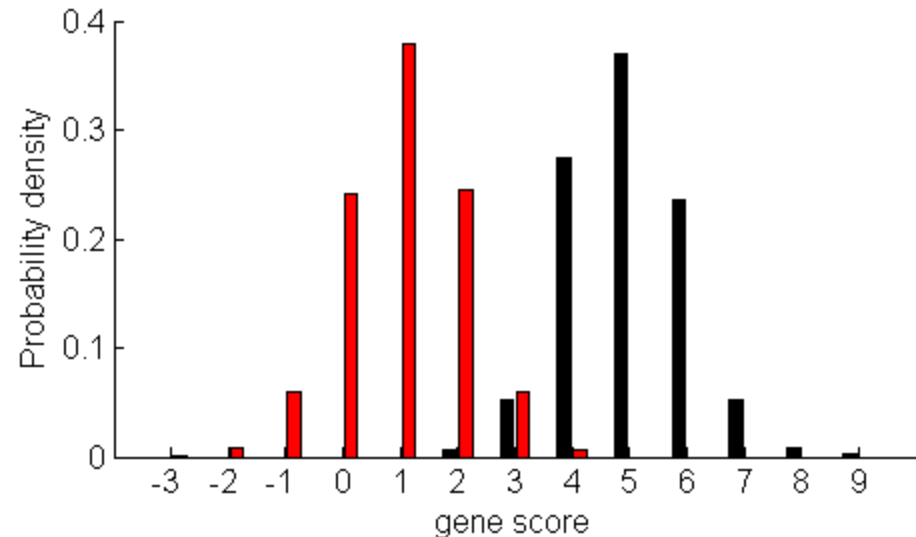
Reminder: ORA using the T-test

P-value = shaded area * 2



$$\begin{aligned} \text{T-statistic} &= \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \\ &= -88.5 \end{aligned}$$

Gene score distributions



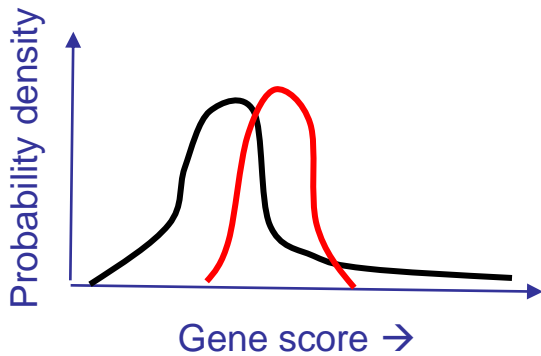
Formal Question: *What is the probability of observing the T-statistic or one more extreme if the means of the two distributions were the same?*

Why can't we use the T-test?

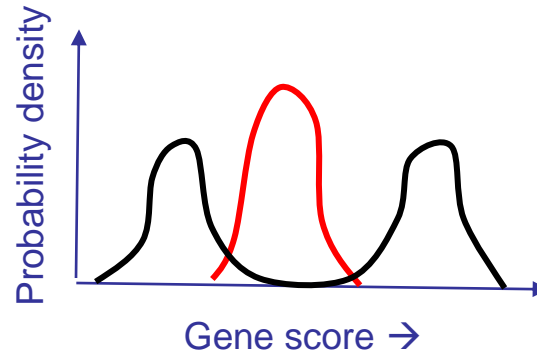
1. Assumes black and red gene score distributions are both approximately Gaussian (i.e. normal)
 - Score distribution assumption is often true for:
 - Log ratios from microarrays
 - Score distribution assumption is rarely true for:
 - Peptide counts, sequence tags (SAGE or NextGen sequencing), transcription factor binding sites hits
2. Tests for significance of difference in means of two distribution but does not test for other differences between distributions.

Examples of inappropriate score distributions for T-tests

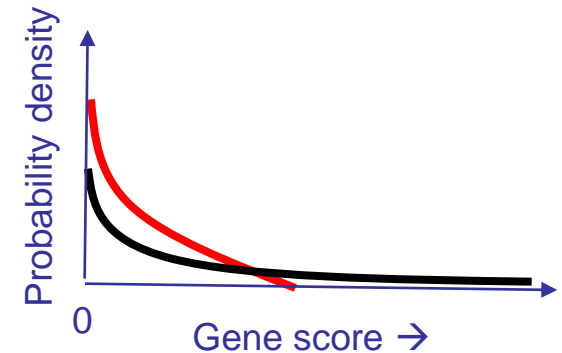
Distributions with gene score outliers, or “heavy-tailed” distributions



Bimodal “two-bumped” distributions.



Gene scores are positive and have increasing density near zero, e.g. sequence counts



Solutions:

- 1) Robust test for difference of medians (WMW)
- 2) Direct test of difference of distributions (K-S)

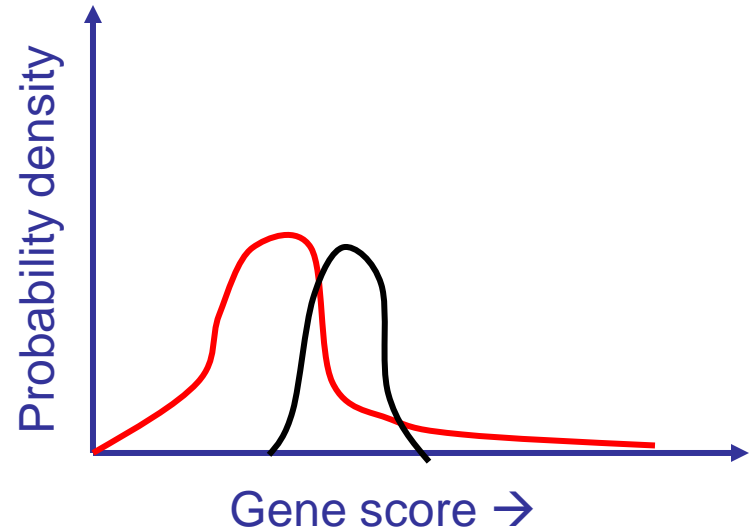
Wilcoxon-Mann-Whitney (WMW) test

aka Mann-Whitney U-test, Wilcoxon rank-sum test

1) Rank gene scores, calculate R_B ,
sum of ranks of black gene scores

			ranks	
	2.1		6.5	1
	5.6		5.6	2
	-1.1		4.5	3
	-2.5		3.2	4
	-0.5		2.1	5
N_2 red gene scores	3.2	3.2	1.7	6
	1.7		0.1	7
	6.5		-1.1	8
	4.5		-2.5	9
	0.1		-0.5	10
N_1 black gene scores				

$R_B = 21$



Formal Question: *Are the medians of the two distributions significantly different?*

Wilcoxon-Mann-Whitney (WMW) test

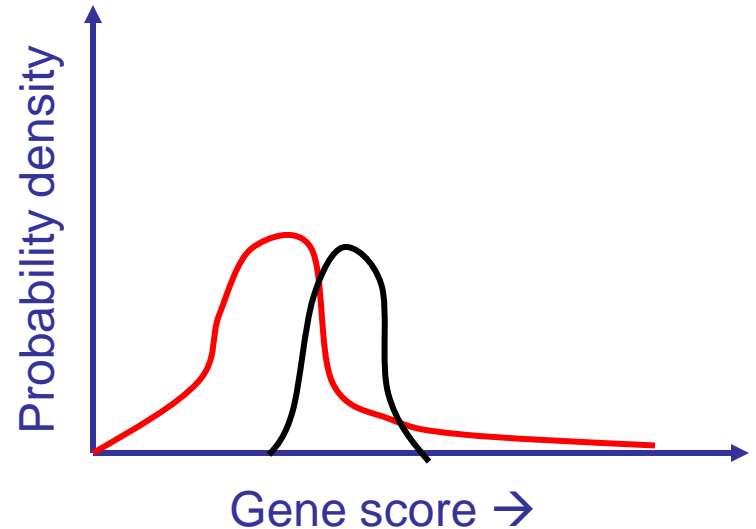
aka Mann-Whitney U-test, Wilcoxon rank-sum test

2) Calculate Z-score:

$$Z = \frac{R_B - N_1 \left(\frac{N_1 + N_2 + 1}{2} \right)}{\sigma_U} = -1.4$$

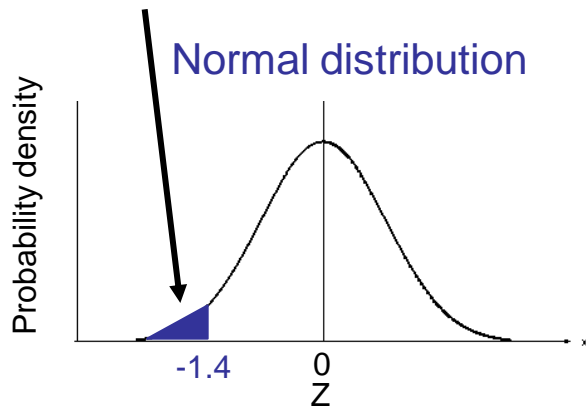
mean rank

$R_B = 21$



3) Calculate P-value:

P-value = shaded area * 2



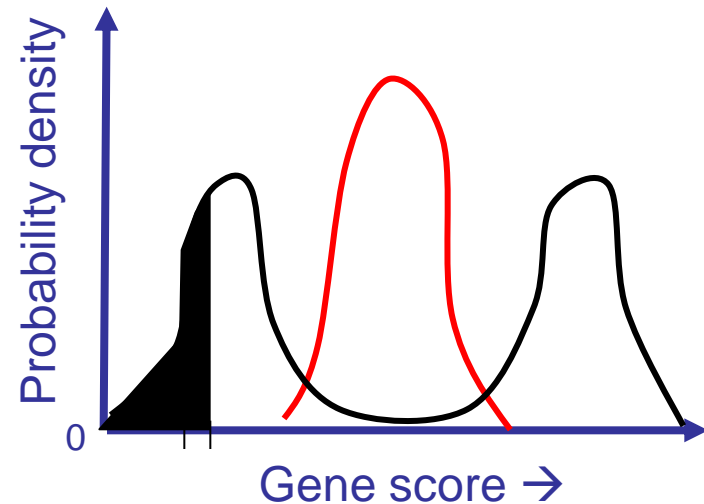
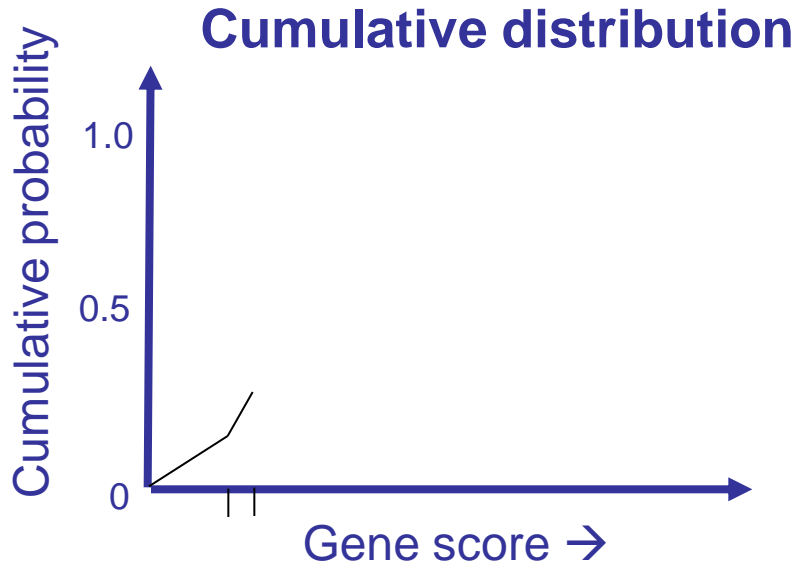
Formal Question: *Are the medians of the two distributions significantly different?*

WMW test details

- Described method is only applicable for large N_1 and N_2 and when there are no tied scores, WMW software uses a tied rank correction
- In most cases the WMW test is simply a T-test applied to the ranks of the gene scores
- WMW test is robust to (a few) outliers
-

$$\sigma_u = \sqrt{N_1 N_2 (N_1 + N_2 + 1) / 12}$$

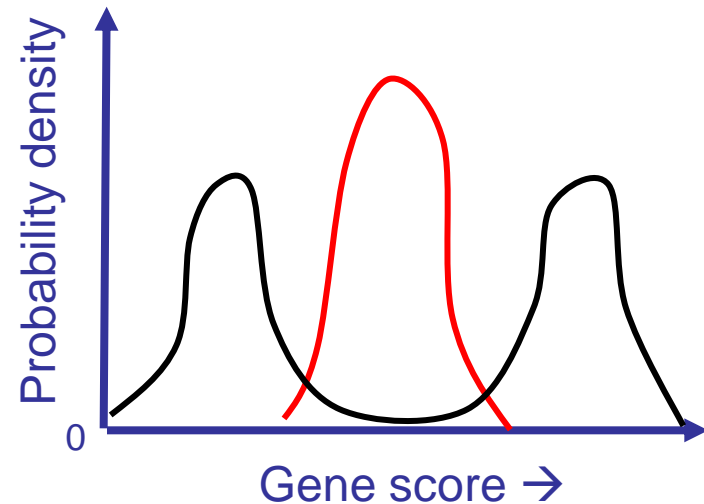
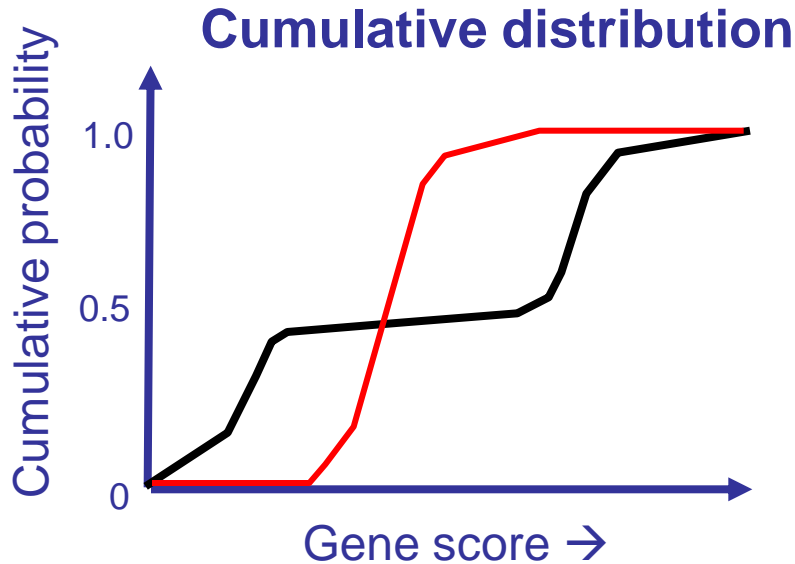
Kolmogorov-Smirnov (K-S) test



1) Calculate cumulative distributions of **red** and black

Question: *Are the red and black distributions significantly different?*

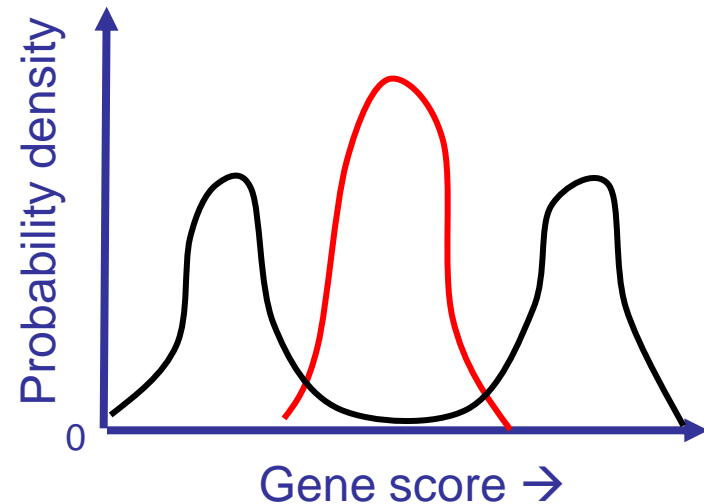
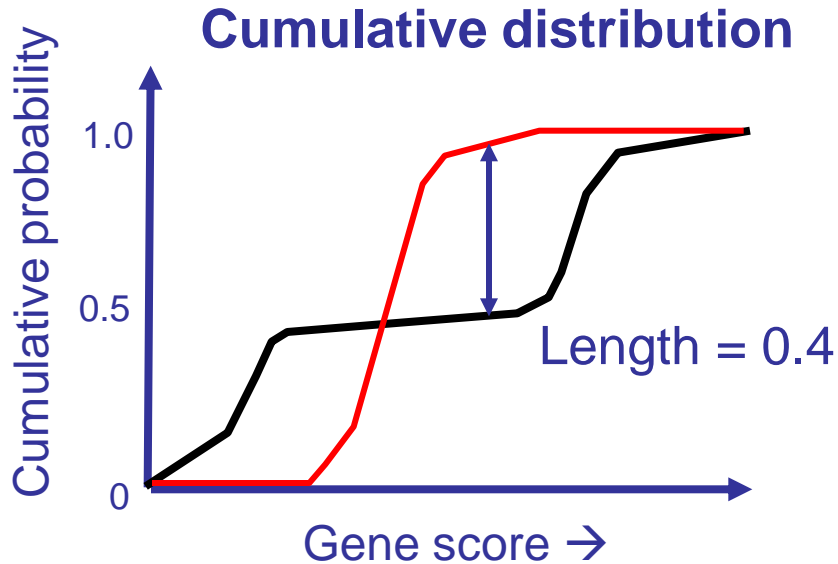
Kolmogorov-Smirnov (K-S) test



1) Calculate cumulative distributions of **red** and black

Question: *Are the red and black distributions significantly different?*

Kolmogorov-Smirnov (K-S) test



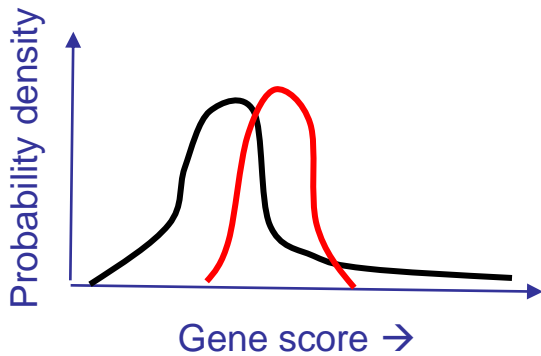
Formal question: *Is the length of largest difference between the “empirical distribution functions” statistically significant?*

WMW and K-S test caveats

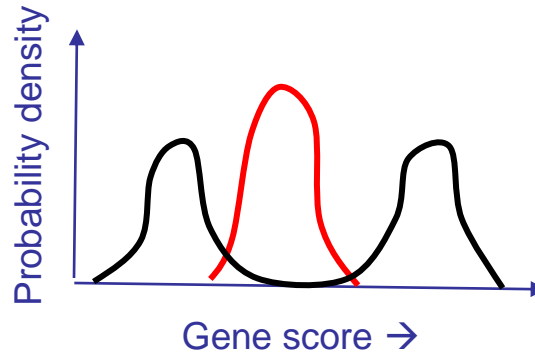
- Neither tests is as sensitive as the T-test, ie they require more data points to detect the same amount of difference, so use the T-test whenever it is valid.
- K-S test and WMW can give you different answers: K-S detects difference of distributions, WMW detects difference of medians
- Rare problem: Tied scores and small # of observations can be a problem for some implementations of the WMW test

Proper tests for different distributions

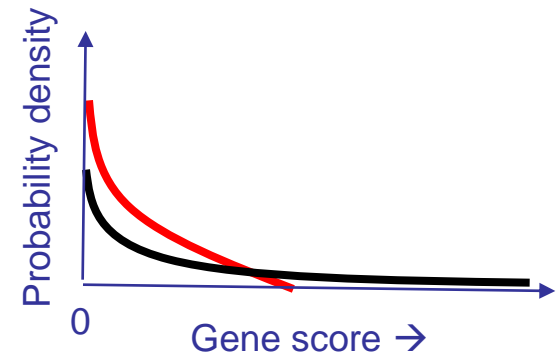
Distributions with gene score outliers, or “heavy-tailed” distributions



Bimodal “two-bumped” distributions.



Gene scores are positive and have increasing density near zero, e.g. sequence counts



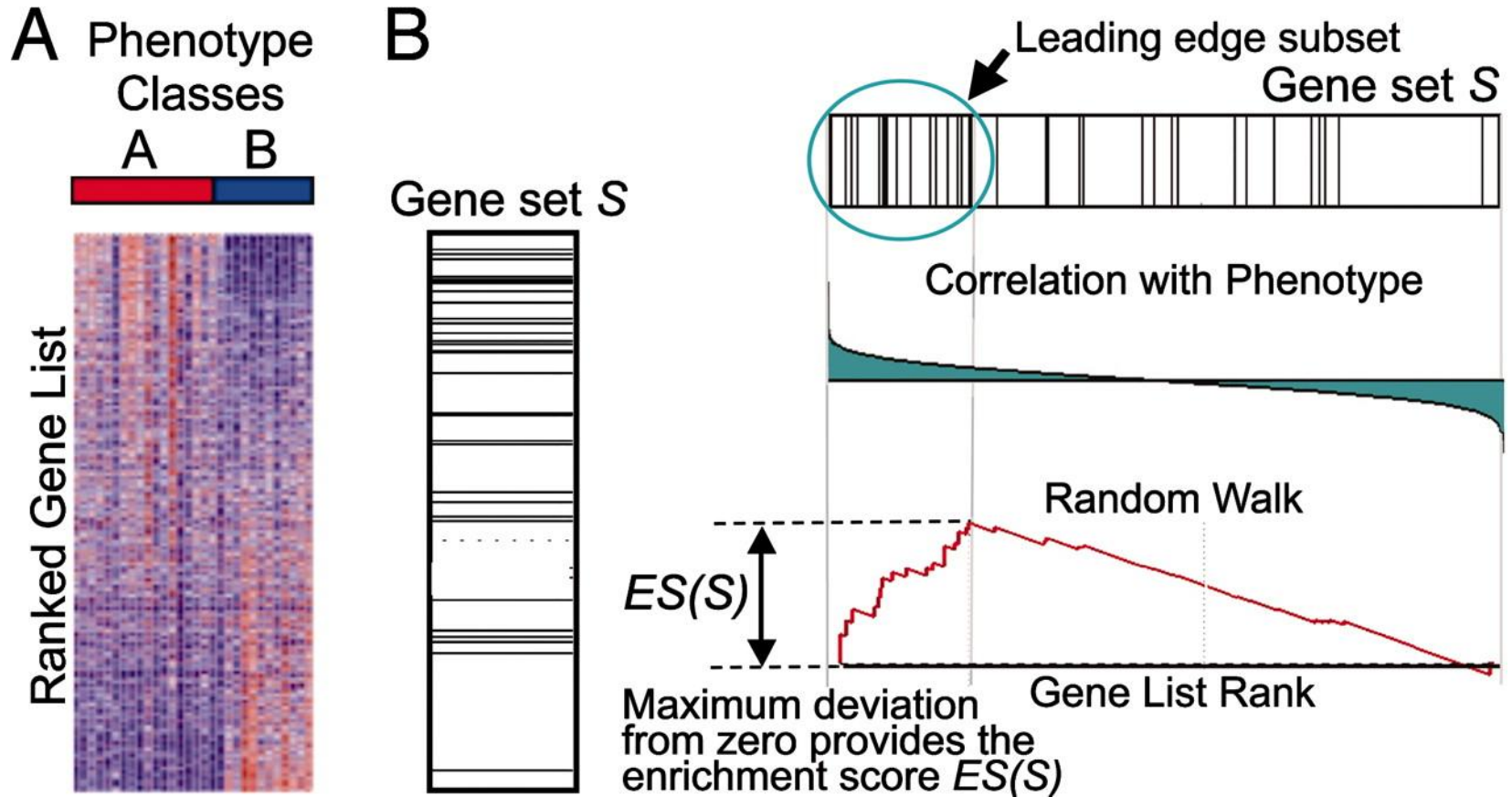
Recommended test:

WMW or K-S

K-S only

WMW or K-S

A GSEA overview illustrating the method



Subramanian A. et.al. PNAS;2005;102:15545-15550

What have we learned?

- T-test is not valid when one or both of the score distributions is not normal,
- If need a “robust” test, or to test for difference of medians use WMW test or GSEA,
- To test for overall difference between two distributions, use K-S test.

Other common tests and distributions

- Chi-squared (contingency table) test
 - Useful if there are >2 values of annotation (e.g. **red genes**, black genes, and **blue genes**)
 - Used as an approximation to Fisher's Exact Test but is inaccurate for small gene lists
- Binomial test
 - Tests if gene scores for **red** and black either come from either N flips of the same coin or different coins.
 - E.g. black genes are “expressed” in, on average, 5 out of 12 conditions and **red genes** are expressed in, on average, 2 out of 12 conditions, is the probability of being expressed significantly different for the black and **red** genes?

Overview

- Theory:
 - Review: What is a P-value? The good ole' T-test.
 - Fisher's Exact Test, the bread and butter of ORA
 - Correcting for multiple testing
 - Enrichment analysis with gene rankings

Break

- Try out an over-representation analysis on gene ranks using GSEA
- GSEA:
 - <http://www.broad.mit.edu/gsea/>
 - <http://www.broadinstitute.org/gsea/downloads.jsp>
 - http://www.broadinstitute.org/gsea/doc/desktop_tutorial.jsp
 - <http://www.broadinstitute.org/gsea/datasets.jsp>

Questions?

Network Visualization/Analysis

Introduction

- Network visualization and analysis using Cytoscape software
- Focus on Cytoscape basics now
- Network analysis using Cytoscape in module 4

<http://cytoscape.org>

Network visualization and analysis

- Pathway comparison
- Literature mining
- Gene Ontology analysis
- Active modules
- Complex detection
- Network motif search

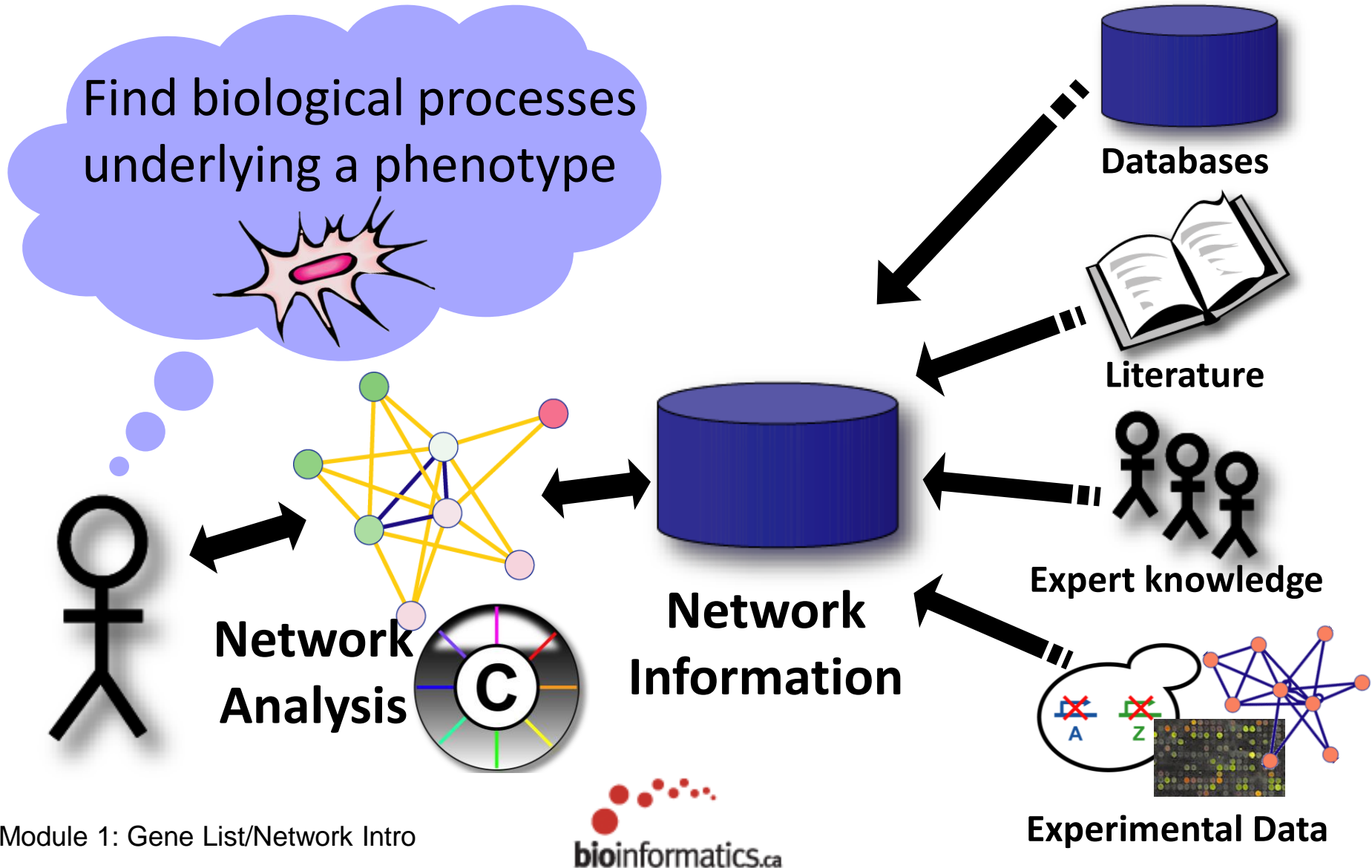
UCSD, ISB, Agilent,
MSKCC, Pasteur, UCSF,
Unilever, UToronto, U
Texas

Network	Nodes	Edges
galFiltered.sif	331(19)	362(35)

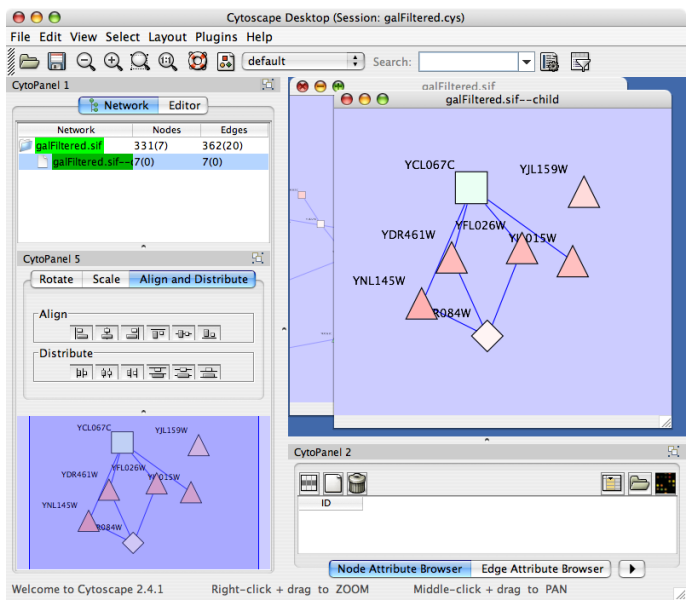
ID	gal1RGexp	gal1RGsig	gal4RGexp	gal4RGsig	gal80Rexp	gal80Rsig
YGL008C	-0.352	1.0007E-5	-0.282	7.1366E-4	-0.573	1.2622E-5
YCL067C	0.169	0.0012873	-0.085	0.11481	0.301	0.0027E-5
YNL145W	-0.764	3.148E-11	-0.098	0.05338	-1.237	1.1916E-5
YMP043W	-0.183	0.0035372	-0.654	4.2514E-6	0.457	2.4112E-5

Welcome to Cytoscape 2.4.0-b1
Right-click + drag to ZOOM
Middle-click + drag to PAN

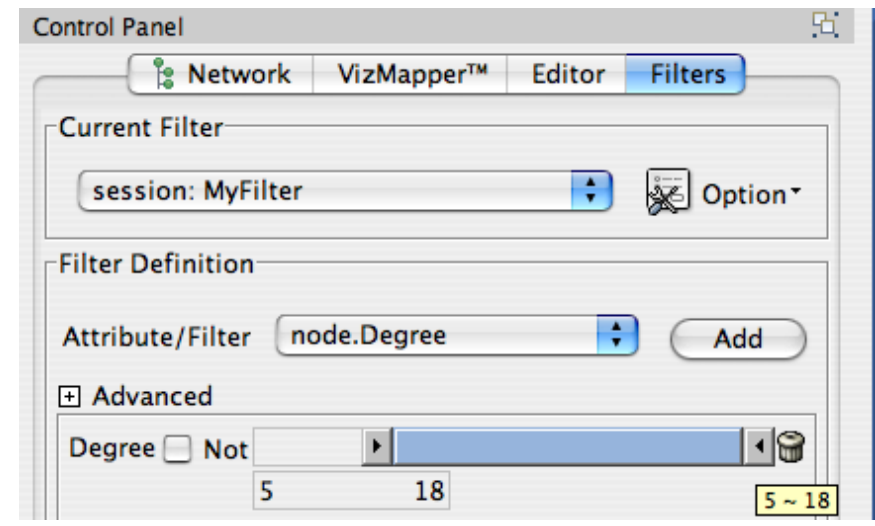
Network Analysis using Cytoscape



Manipulate Networks

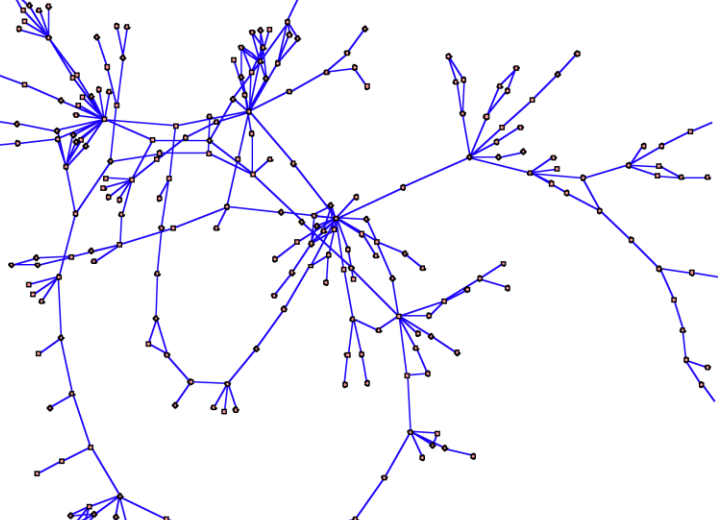


Filter/Query

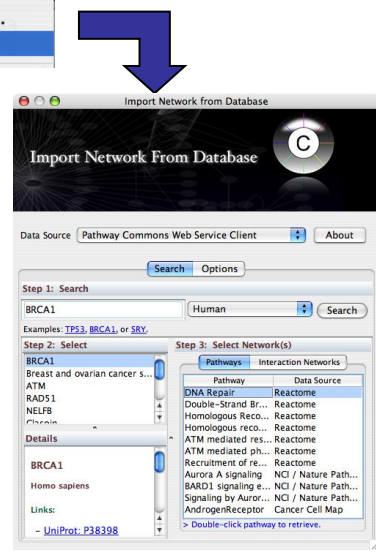
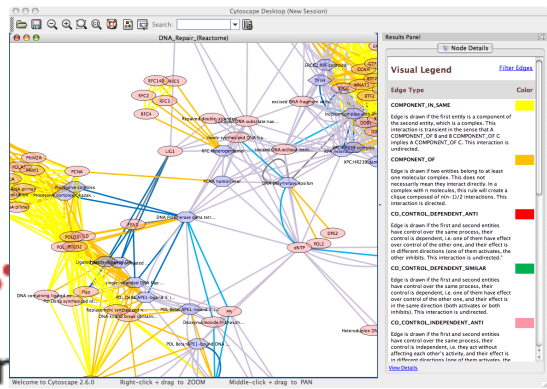
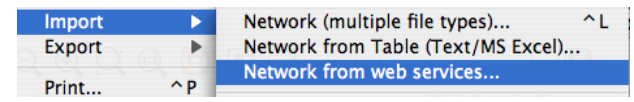


Interaction Database Search

Automatic Layout

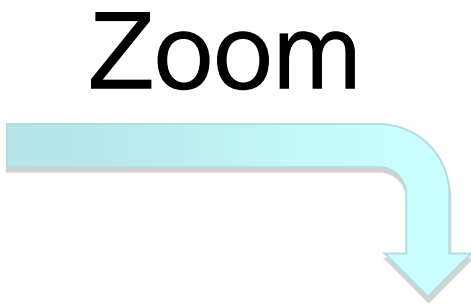
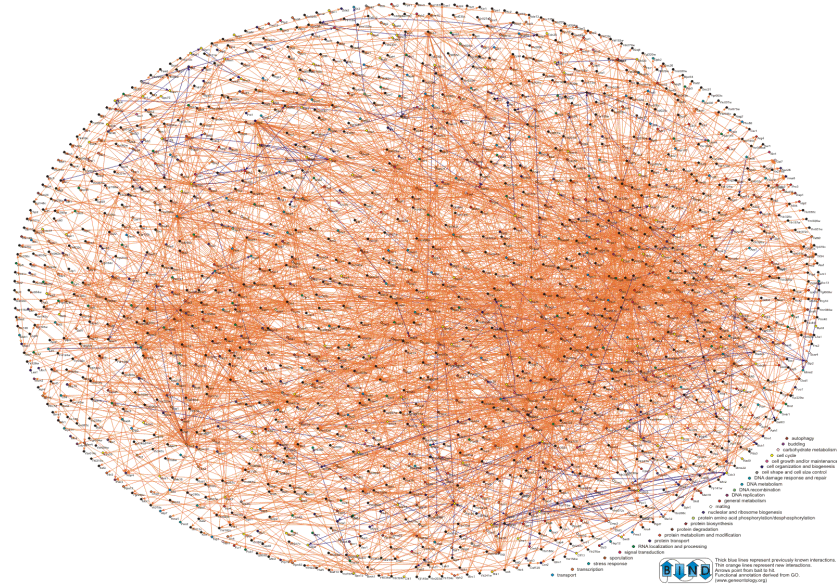


Module 1: Gene List/Network Intro



Overview

Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry

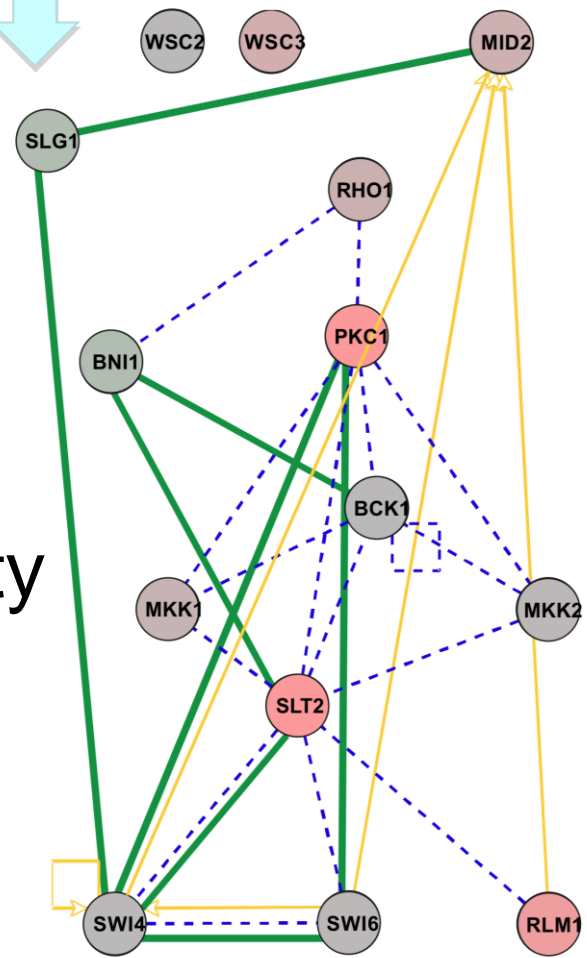


Zoom

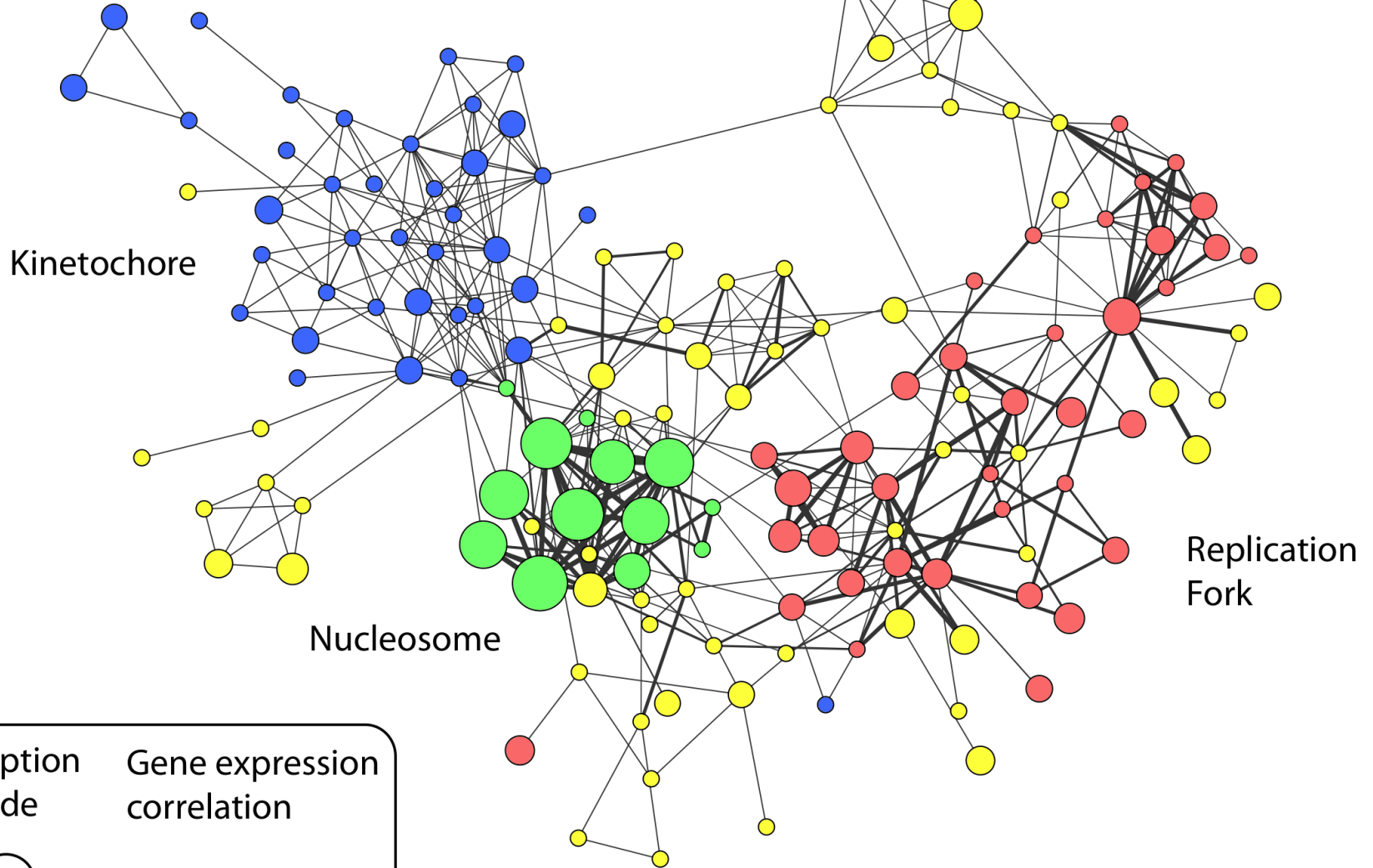
Focus

PKC Cell Wall Integrity

- Synthetic Lethal
- Transcription Factor Regulation
- Protein-Protein Interaction
- Up Regulated Gene Expression
- Down Regulated Gene Expression



Visualize multiple data types on a network



Transcription
amplitude

Gene expression
correlation



low high



low high

Control: node/edge size, shape, color...

Active Community

<http://www.cytoscape.org>

- Help

- 8 tutorials, >10 case studies
- Mailing lists for discussion
- Documentation, data sets

Cline MS et al. Integration of biological networks and gene expression data using Cytoscape Nat Protoc. 2007;2(10):2366-82

- Annual Conference: Houston Nov 6-9, 2009
- 10,000s users, 2500 downloads/month
- >40 Plugins Extend Functionality
 - Build your own, requires programming

What Have We Learned?

- Cytoscape is a useful, free software tool for network visualization and analysis
- Provides basic network manipulation features
- Plugins are available to extend the functionality

Cytoscape Demo

Version 2.6

www.cytoscape.org

FYI

Desktop

Control Panel

Network

Network	Nodes	Edges
galFiltered.s	331(4)	362(0)

Network manager

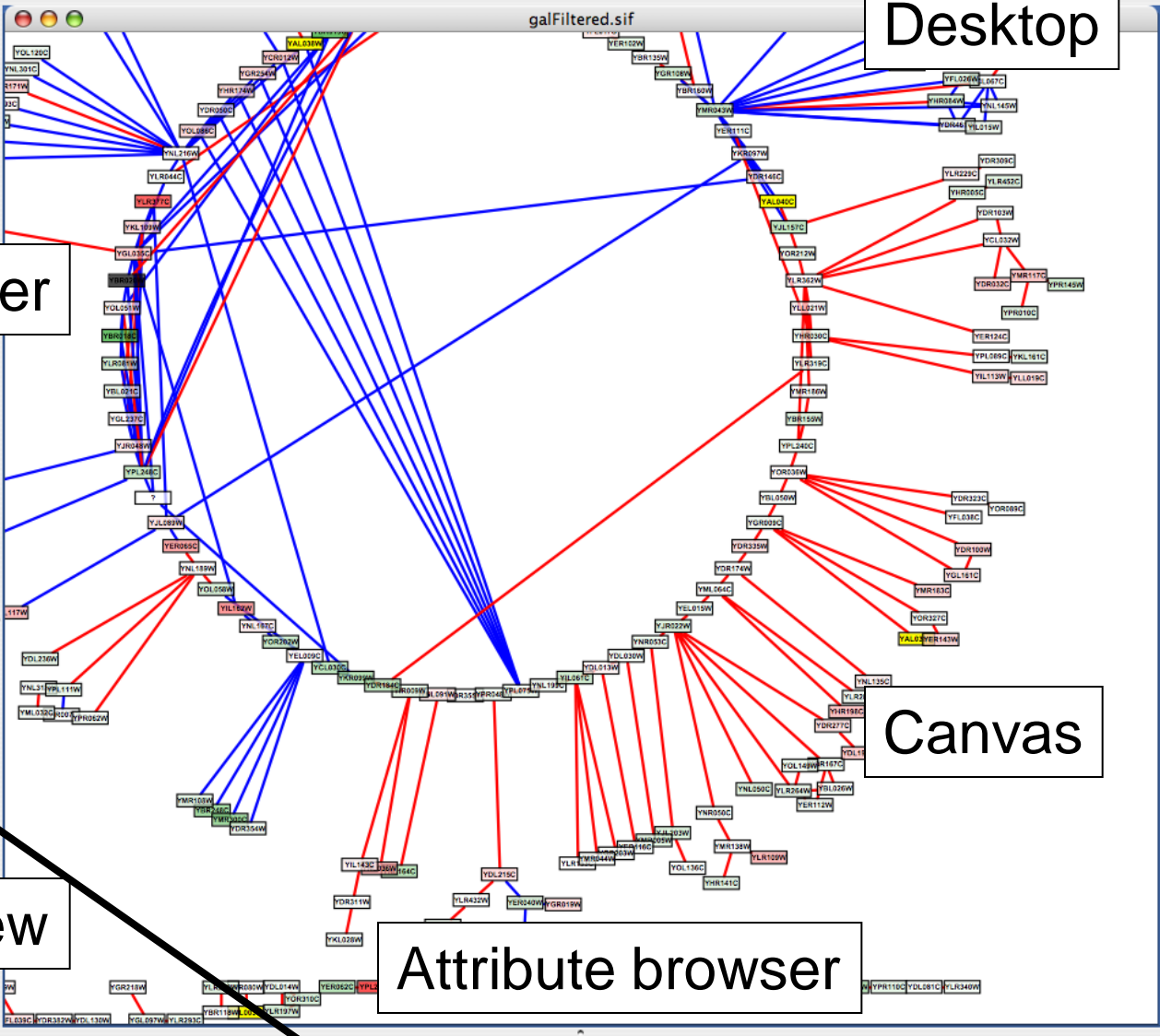
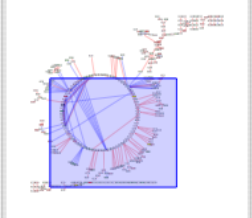
CytoPanels

Network overview

Attribute browser

Canvas

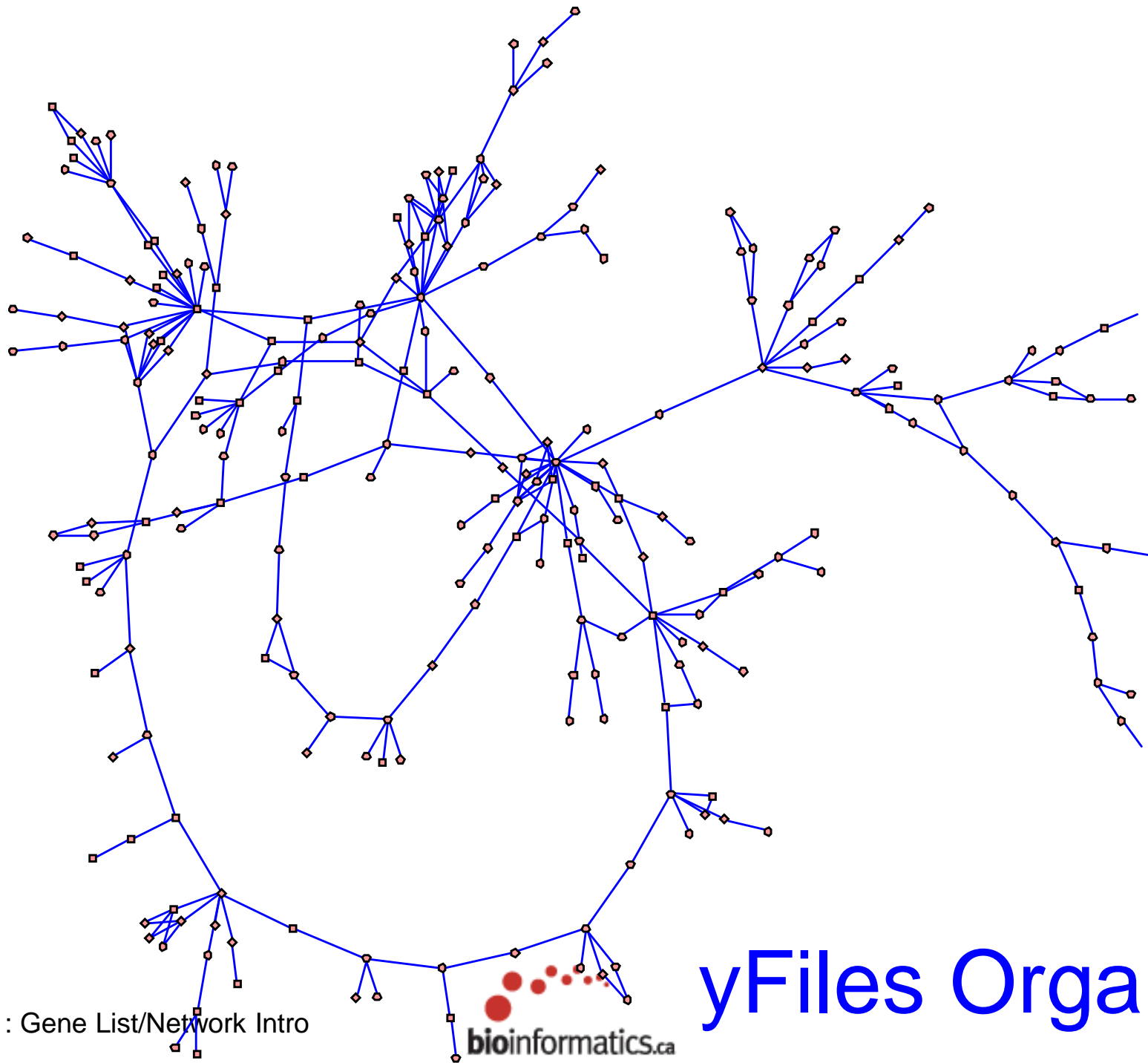
Module 1:



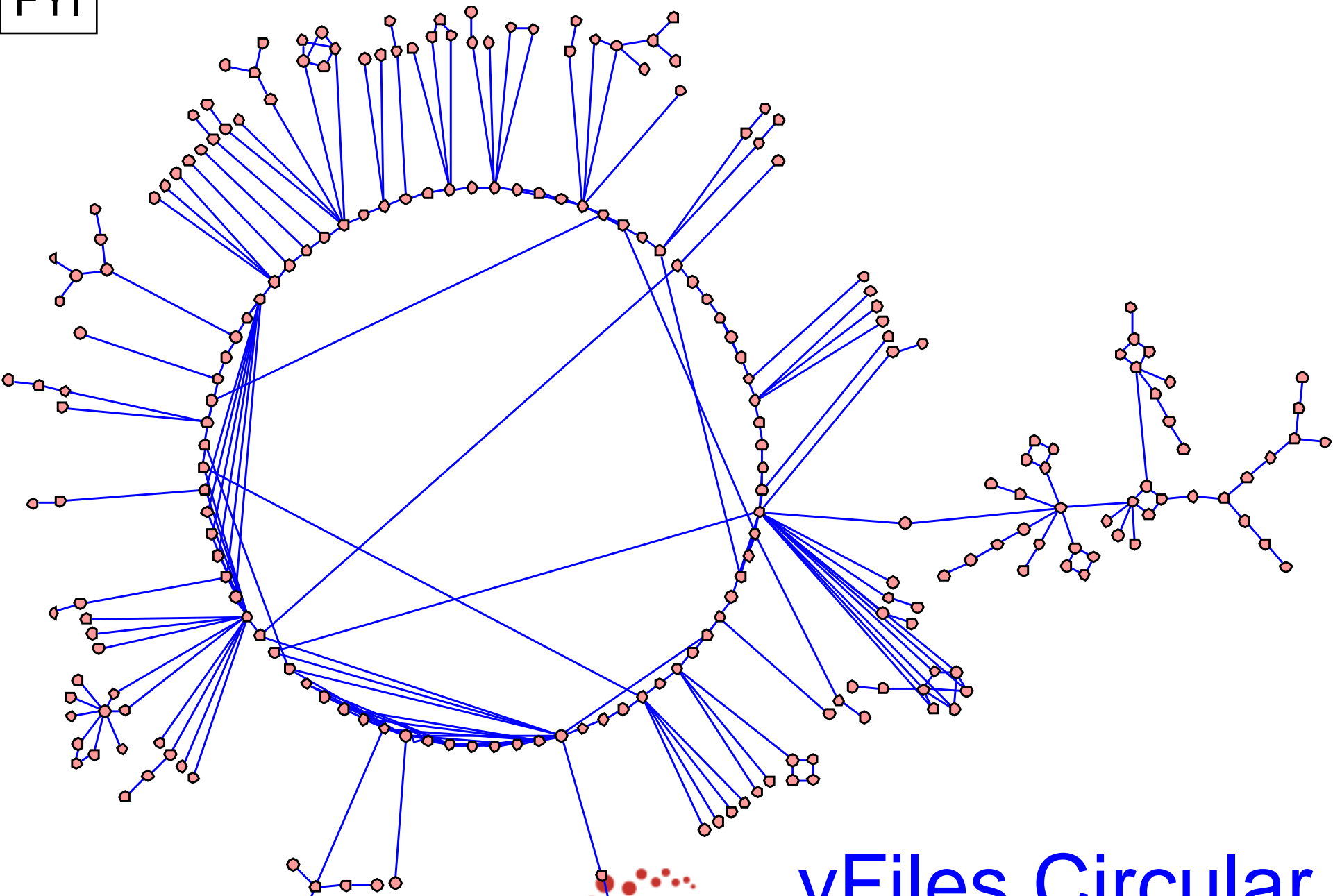
Data Panel

ID	Degree	gal1RGexp	gal1RGsig	gal4RGexp	gal4RGsig	gal80Rexp	gal80Rsig
YAL038W	3	-0.652	1.3173E-10	0.123	0.10377	-0.453	1.5489E-7
YAL003W	2	-0.157	0.0018696	-0.2	6.2814E-4	-0.146	0.013062
YAL030W	2	-0.05	0.18782	0.027	0.55374	0.0030	0.95396

Node Attribute Browser Edge Attribute Browser Network Attribute Browser



FYI



Module 1: Gene List/Network Intro



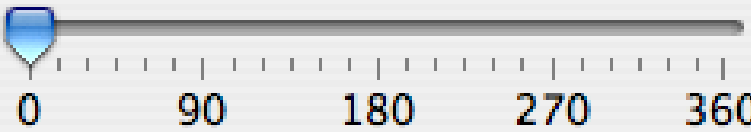
yFiles Circular

Network Layout

- 15 algorithms available through plugins
- Demo: Move, zoom/pan, rotate, scale, align

Rotate | Scale | Align and Distribute

Rotate in Degrees:




0 90 180 270 360


Rotate Selected Nodes Only

Rotate | Scale | Align and Distribute

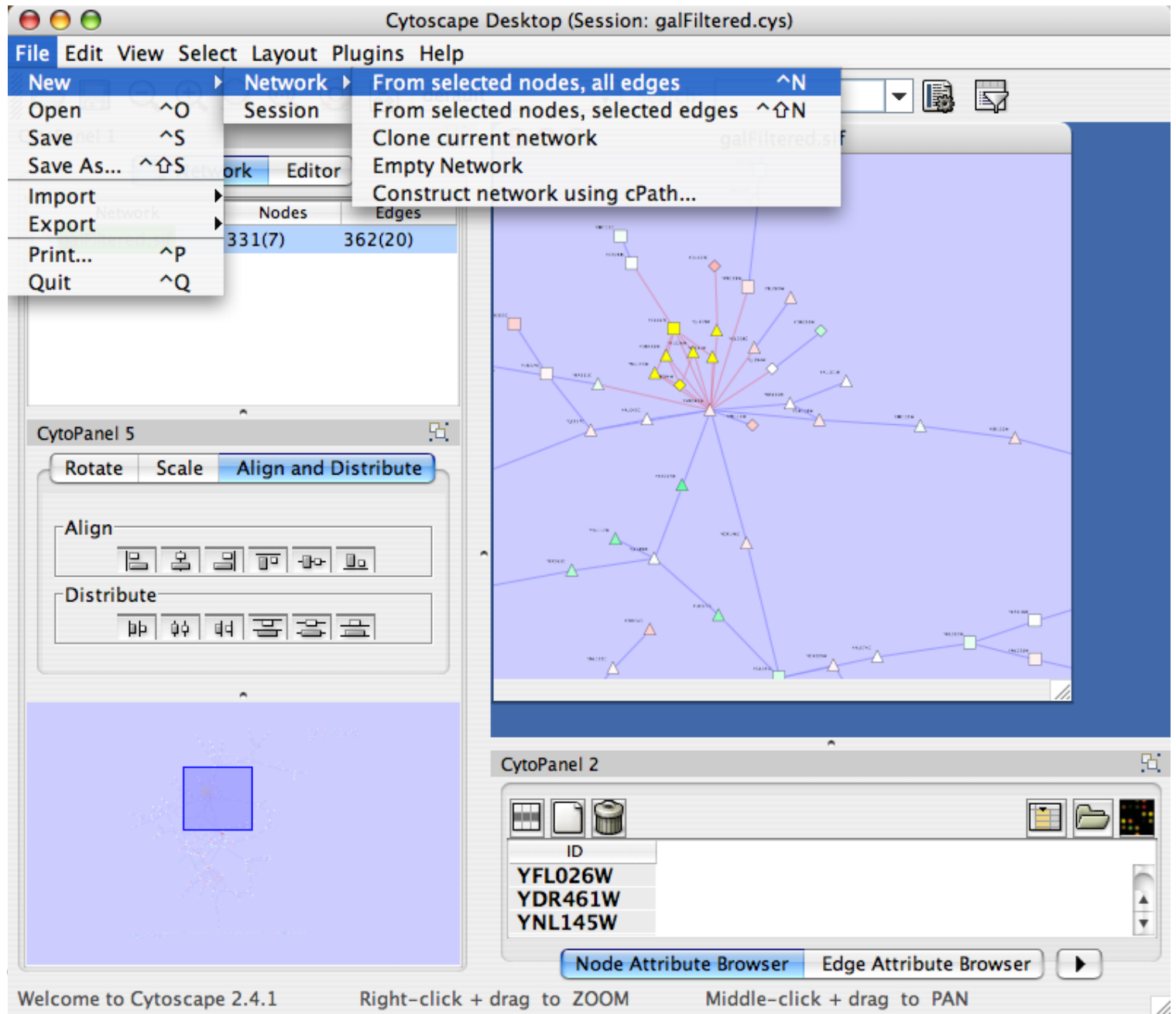
Align



Distribute



Create Subnetwork



Module 1:

Create Subnetwork

The screenshot displays the Cytoscape Desktop interface with the following components:

- Menu Bar:** File, Edit, View, Select, Layout, Plugins, Help
- Toolbar:** Includes icons for file operations, zooming, and search.
- CytoPanel 1 (Network Editor):**

Network	Nodes	Edges
galFiltered.sif	331(7)	362(20)
galFiltered.sif--	7(0)	7(0)
- CytoPanel 5 (Align and Distribute):** Contains tools for rotating, scaling, and aligning/distributing nodes.
- Network Visualization:** A graph with nodes YCL067C (green square), YJL159W (pink triangle), YDR461W (pink triangle), YFL026W (pink triangle), YNL145W (pink triangle), YW015W (pink triangle), and R084W (pink triangle) connected by blue edges. A white diamond node is also present at the bottom.
- CytoPanel 2 (Node Attribute Browser):** Shows a table with an 'ID' column.
- Footer:** Welcome to Cytoscape 2.4.1, Right-click + drag to ZOOM, Middle-click + drag to PAN

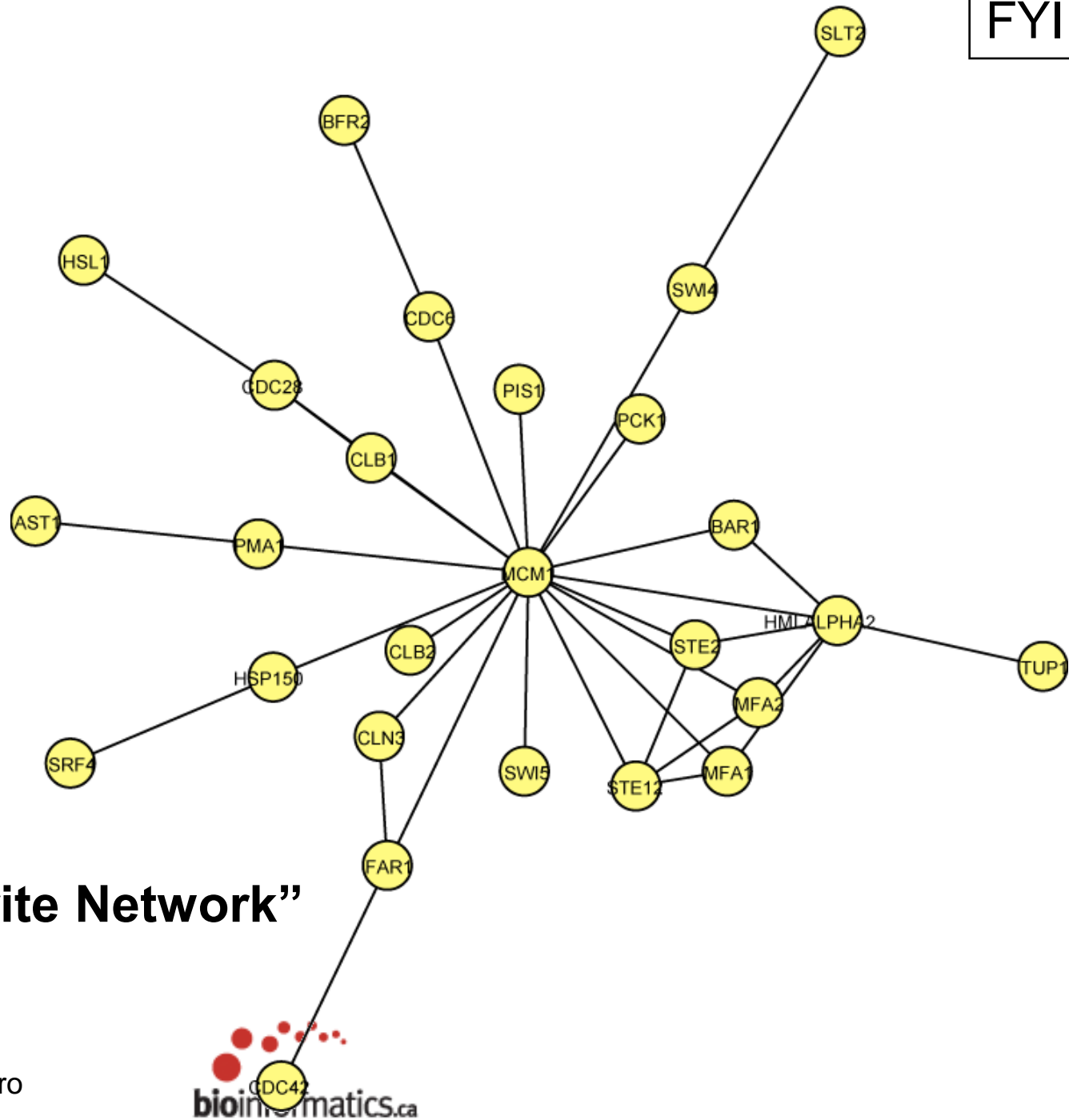
Module 1:

Visual Style

- Customized views of experimental data in a network context
- Network has node and edge attributes
 - E.g. expression data, GO function, interaction type
- Mapped to visual attributes
 - E.g. node/edge size, shape, colour...
- E.g. Visualize gene expression data as node colour gradient on the network

Visual Style

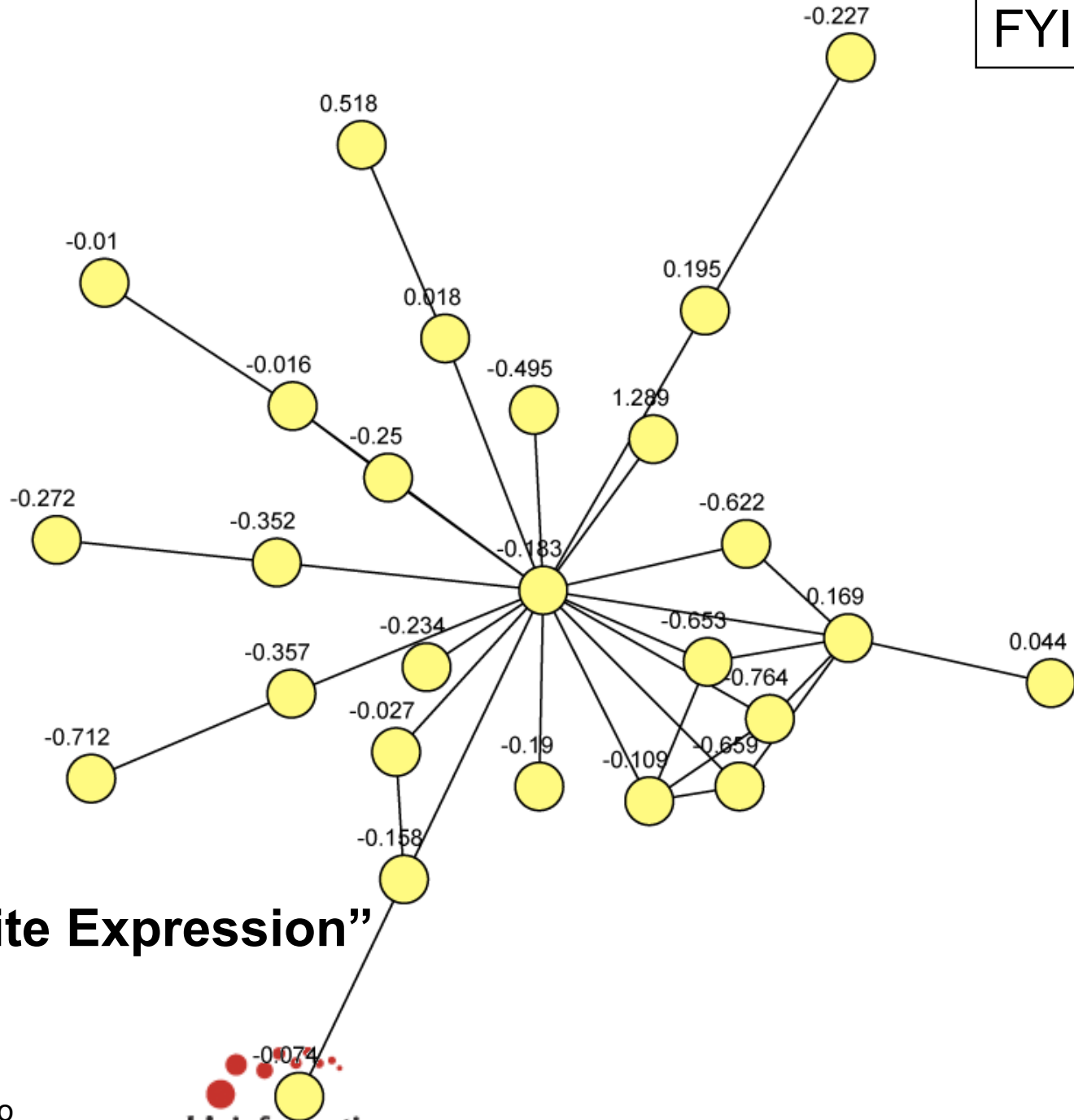
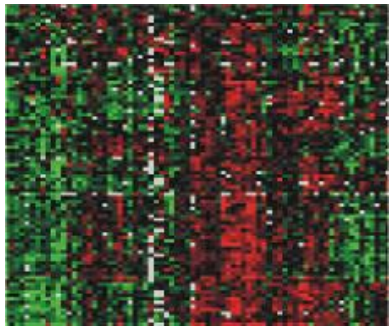
FYI



Load “Your Favorite Network”

Visual Style

FYI



Load “Your Favorite Expression” Dataset

Visual Style

Cytoscape Desktop (Session: galFiltered.cys)

Control Panel

Network VizMapper™ Editor Filters


Current Visual Style: Sample3

Defaults: Source -> Target

Visual Mapping Browser

- Edge Visual Mapping
 - Edge Color: interaction
- Node Visual Mapping
 - Node Tooltip: gal4RGexp
 - Node Label: ID
 - Node Color: gal4RGexp

Mapping Type: Continuous Mapping

Graphical View:  -2.41 to 1.22

Unused Properties: Node Border Color, Node Shape, Node Width, Node Height

galFiltered.sif

Gradient Editor for Node Color

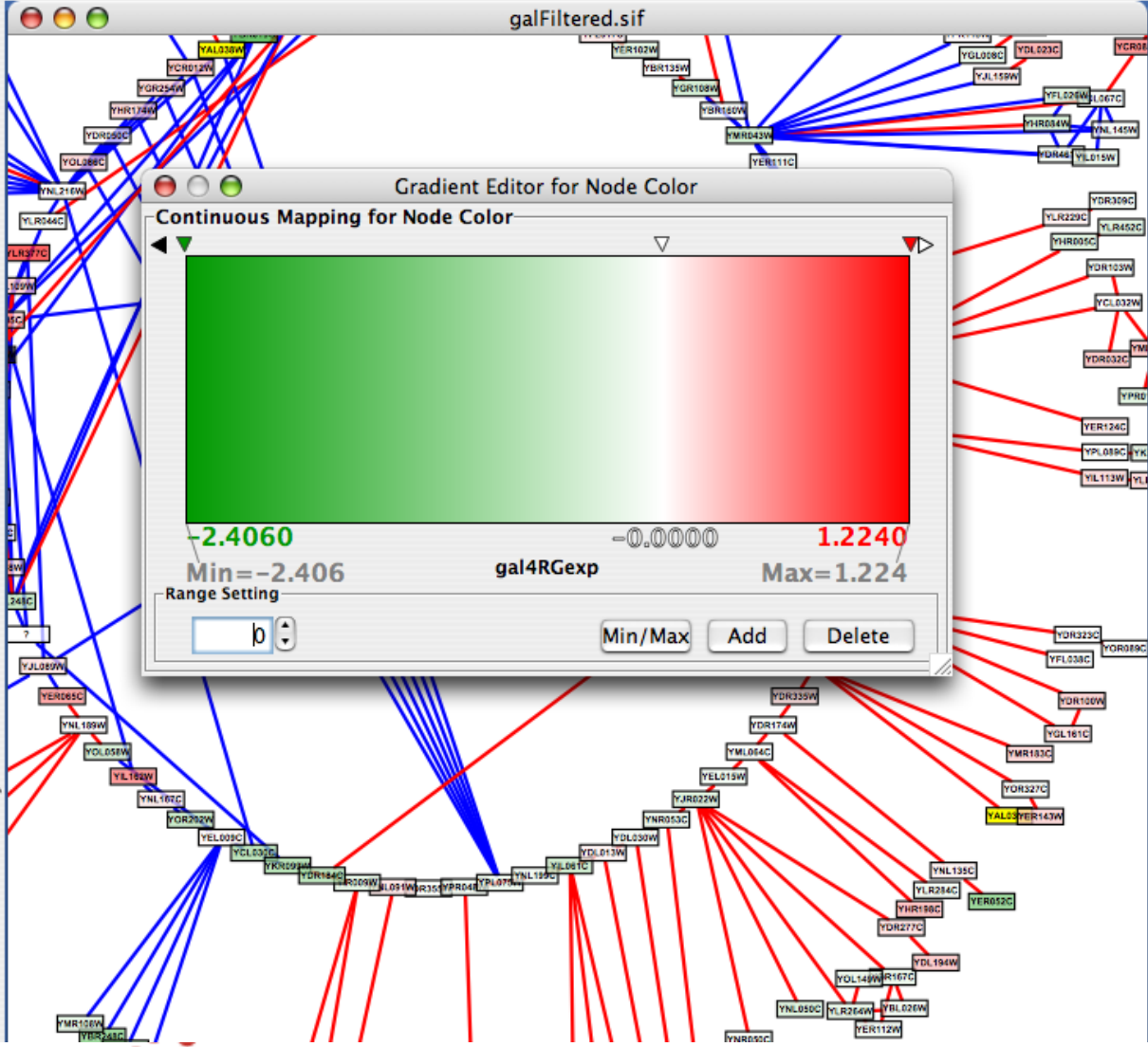
Continuous Mapping for Node Color

Min = -2.406 Max = 1.224

gal4RGexp

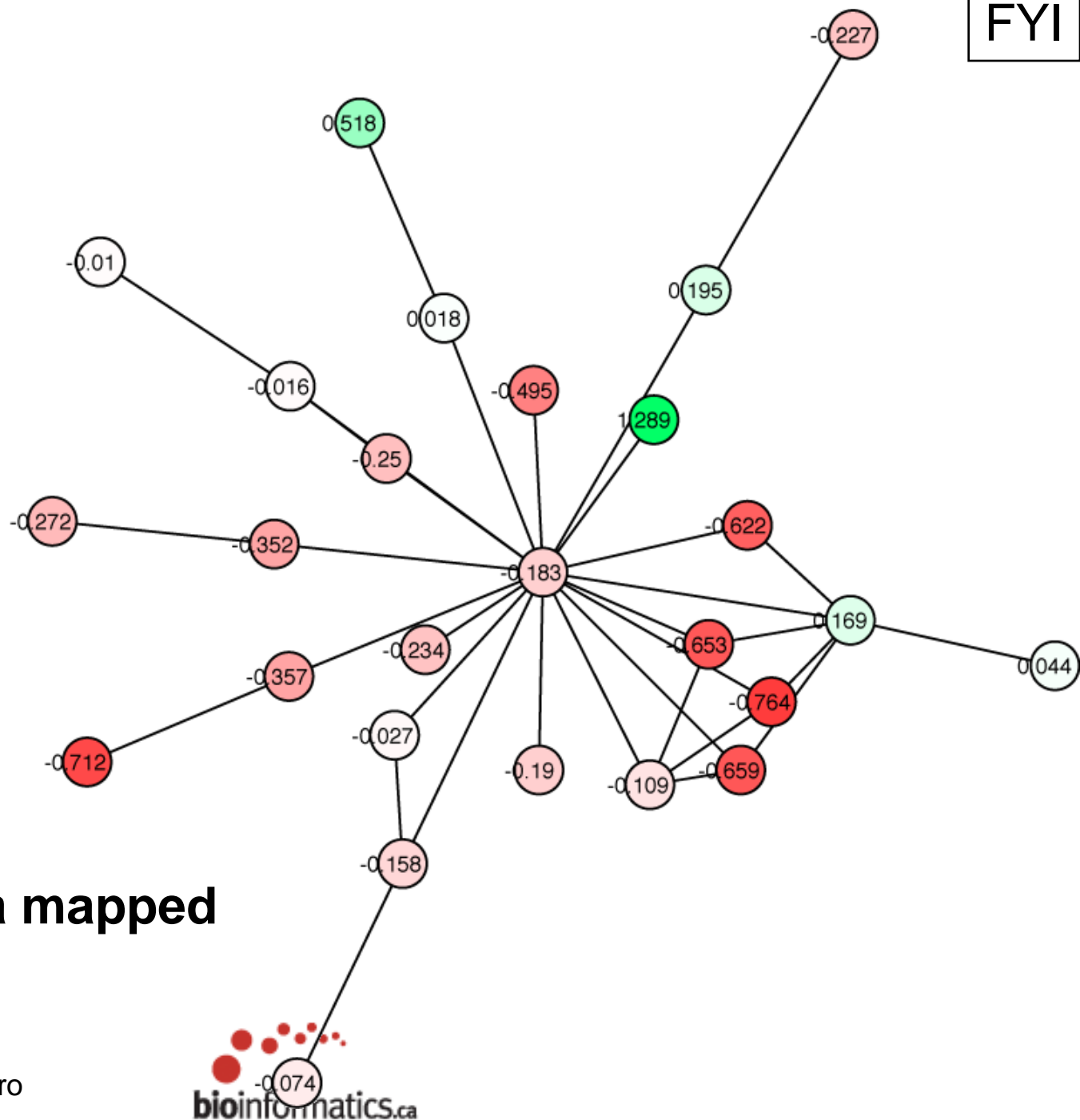
Range Setting:

Buttons: Min/Max, Add, Delete



Visual Style

FYI



Network Filtering

Cytoscape Desktop (Session: galFiltered.cys)

Control Panel

- Network
- VizMapper™
- Editor
- Filters

Current Filter: session: MyFilter

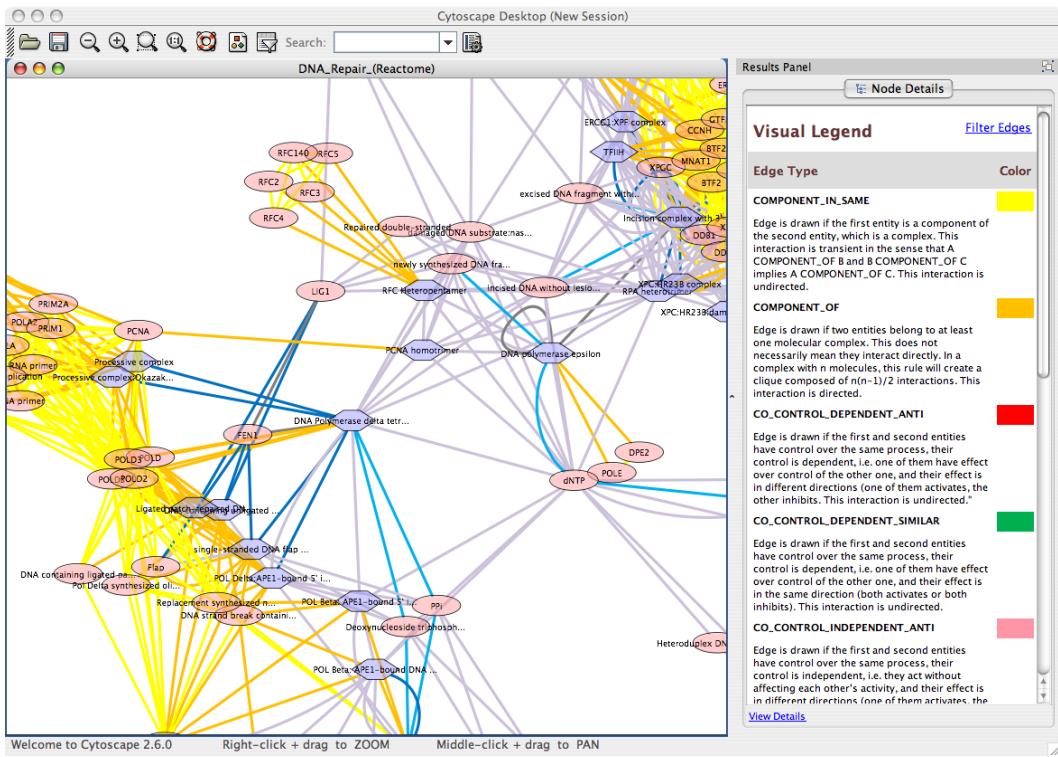
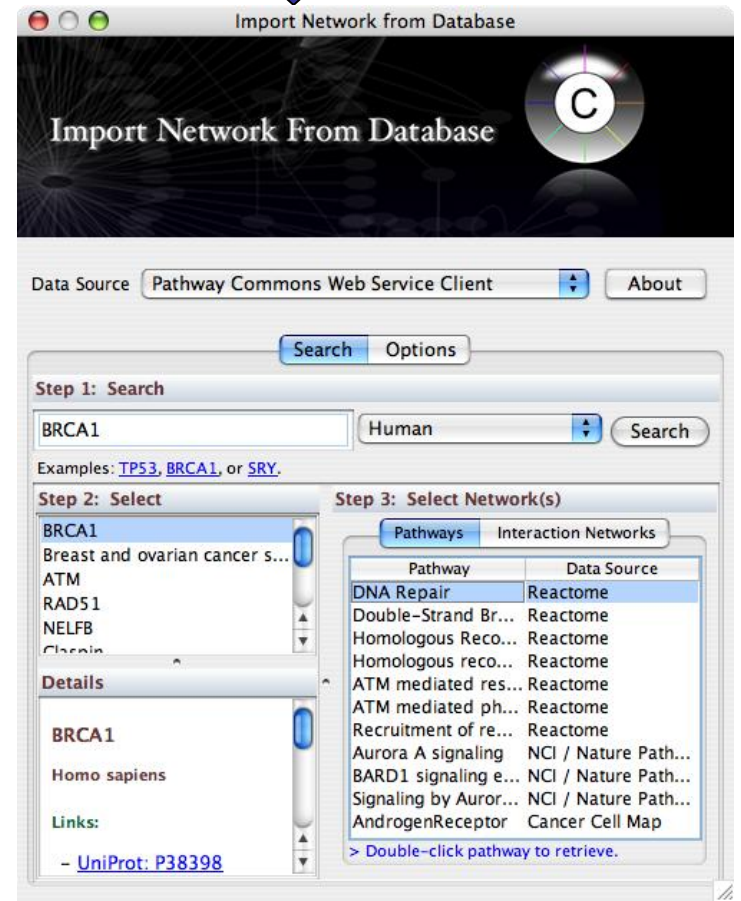
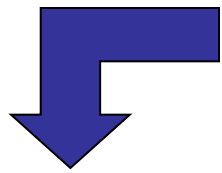
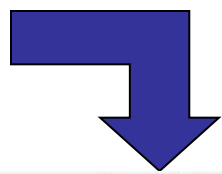
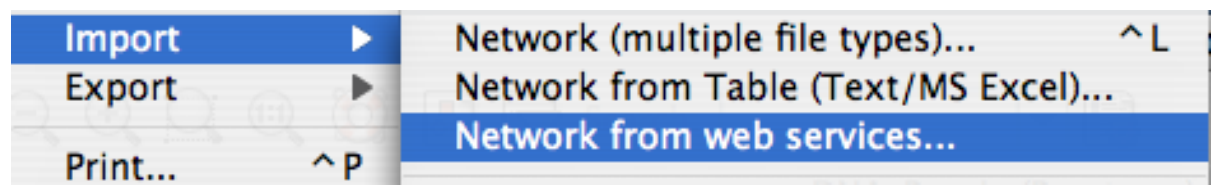
Filter Definition: Attribute/Filter: node.Degree

Advanced: Degree Not 5 ~ 18

The network graph displays a complex web of interactions between various yeast genes. Nodes are represented by small rectangular boxes with their unique identifiers. The graph is filtered based on node degree, with nodes having a degree between 5 and 18 highlighted in yellow. Edges connecting these nodes are colored blue, while other edges are red. The layout is a force-directed graph, showing clusters and hubs of high connectivity.

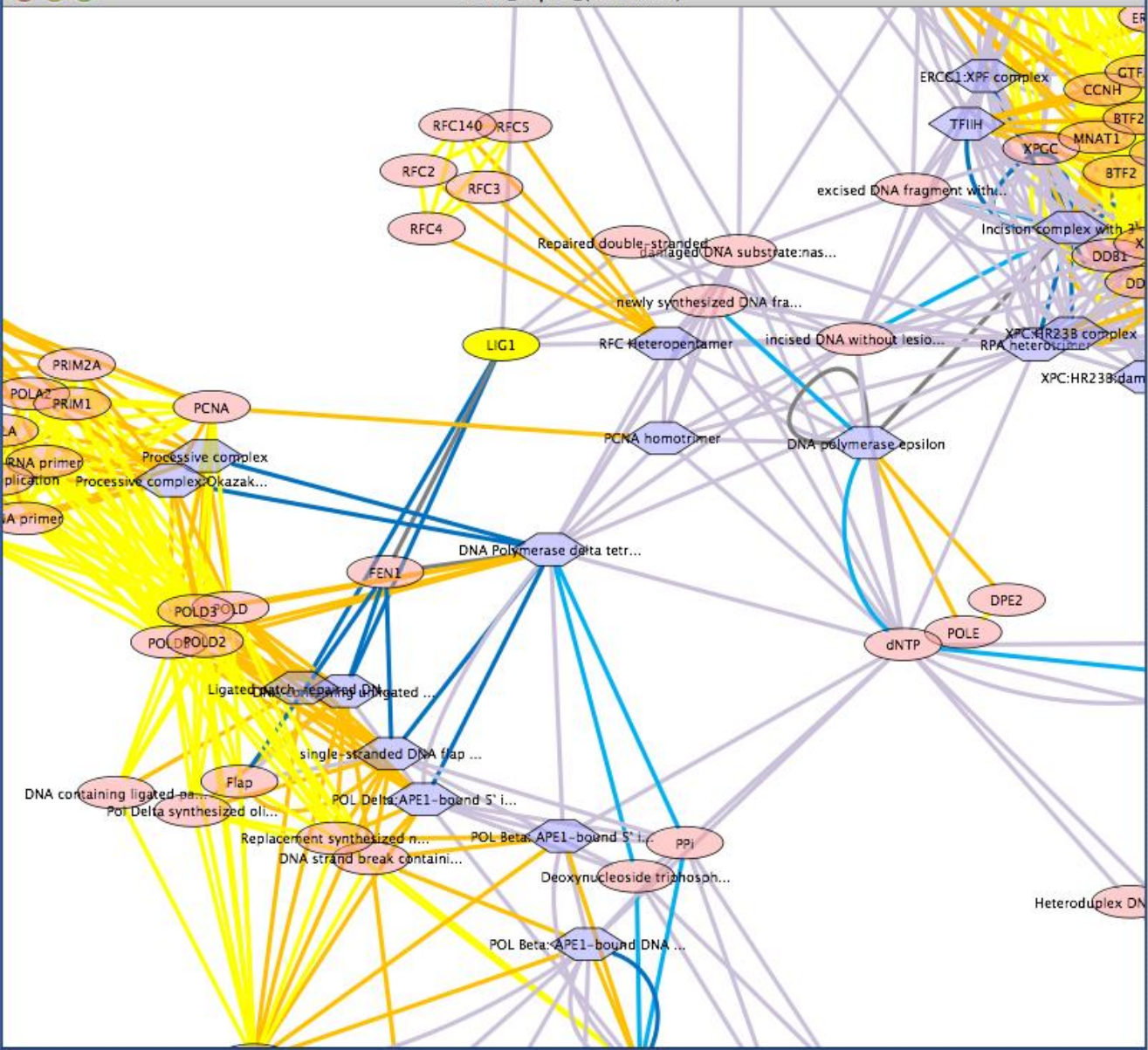
FYI

Interaction Database Search





DNA_Repair_(Reactome)



Results Panel

Node Details

LIG1

Protein

Homo sapiens

[Pathway Commons: 6311](#)

Synonyms:

- LIG1

Links:

- [UNIPROT: P18858](#)
- [UNIPROT: Q32P23](#)
- [REF_SEQ: NP_000225](#)
- [Search iHOP](#)

[Visual Legend](#)

